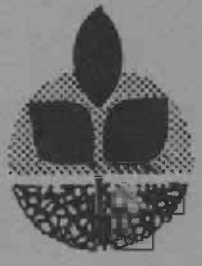

# A Proposed "Partitioned" Ridge Regression Procedure for Estimating Models Unsuited for Regular Ridge Regressions



Special Report 439
June 1975

Agricultural Experiment Station
Oregon State University, Corvallis

ABSTRACT

Earlier papers by Brown [1973] and Brown and Beattie [1975] showed, for the common case of positively interrelated explanatory variables, that the ridge regression estimator of Hoerl and Kennard [1970] gave much larger bias and mean square error (MSE) for models with unlike signs. To avoid this excessive bias and MSE in such cases, a method for partitioning ridge estimation into compatible blocks is proposed in this paper. The expected values and variances of the "partitioned" ridge estimates are derived, and the procedure is applied to a simple experimental model, giving partitioned ridge estimates that were two to three times more accurate in terms of MSE than OLS or regular ridge regression. The partitioned ridge method was also applied to a nonorthogonal subset of data from an empirical production study. Both regular and partitioned ridge regression estimates had about one-half the MSE of OLS for certain ranges of k values. However, the partitioned ridge estimates were superior to the regular ridge estimates in that a lower MSE was obtained over a wider range of k values. Also, a much larger proportion of the partitioned MSE was composed of easily measured variance, as compared to a high percent of bias squared component for the MSE of the regular ridge regression estimator. These results indicate that the partitioned approach should substantially increase the usefulness of biased linear estimation for many regression problems.

AUTHOR: William G. Brown, Professor, Department of Agricultural and Resource Economics, Oregon State University, Corvallis.

ACKNOWLEDGEMENT: Research conducted under Oregon Agricultural Experiment Station Project 128.

The author is deeply indebted to Don A. Pierce for his exceedingly valuable ideas and suggestions, generously contributed at all stages of this research. The author also greatly appreciates the use of David Fawcett's ridge regression computer program, *RIDGE, used in this research. Constructive reviews of this paper were received from Albert Halter, Timothy Hammonds, and John Trierweiler. However, any errors or deficiencies are the sole responsibility of the author.

CONTENTS

# A PROPOSED "PARTITIONED" RIDGE REGRESSION PROCEDURE FOR ESTIMATING MODELS UNSUITED FOR REGULAR RIDGE REGRESSION

William G. Brown

The concept of biased linear estimation received much well-deserved attention from the papers of Hoerl and Kennard [1970, a, b]. They analyzed the "ridge regression" estimator which, although biased, is very effective in reducing the variance of parameter estimates of the general linear regression model fitted to nonorthogonal data. Following Hoerl and Kennard [1970a], assume the linear model,

$$Y = X\beta + u, \tag{1}$$

where Y is $n \times 1$, fixed X is $n \times p$, $\beta$ is $p \times 1$, u is $n \times 1$, $Eu = 0$, and $Euu' = \sigma^2 I$. The ridge estimator, $\hat{\beta}^*$, is defined as

$$\hat{\beta}^* = (X'X + kI)^{-1} X'Y, \tag{2}$$

where X'X represents the correlation matrix of explanatory variables and k denotes a small positive increment.

Although a very small variance can be achieved by increasing the level of k, the bias can become excessive for certain situations. "Excessive" bias from ridge regression can be expected for the fairly common situation of positive interrelationship among the explanatory variables <u>if</u> the true regression parameters are of unlike sign. This result follows from a theorem by Brown [1973] and proven more compactly by Brown and Beattie [1975] following a helpful suggestion from T. D. Wallace. The theorem states that the bias of the ridge estimate of the $j^{\underline{th}}$ standardized regression coefficient can be expressed as

$$E(\hat{\beta}_j^*) - \beta_j = \frac{k\, c_{jj}}{|X'X + kI|} \sum_{i=1}^{p} \hat{b}_{ji}^* \beta_i, \text{ where } \hat{b}_{ji}^* = -1.0 \text{ if } i = j, \tag{3}$$

and if $i \neq j$, $\hat{b}_{ji}^*$ denotes the ridge estimate of the $i^{\underline{th}}$ regression coefficient of

the model where $X_j$ has been regressed on the remaining (p-1) explanatory variables. ($\beta_i$ and $\beta_j$ represent the true regression coefficients in the original model, $Y = X\beta + u$, and $c_{jj}$ is the minor formed by deleting the $j^{\text{th}}$ row and column from $(X'X + kI)$.)

The above theorem is helpful in deciding whether or not to use ridge regression, since one usually obtains small bias and mean square error (MSE) only if (for positively interrelated explanatory variables) the true $\beta$ values all have the same sign, such as for a Cobb-Douglas production function. However, this condition is obviously restrictive, since many economic models have positively interrelated explanatory variables, but with some positive and some negative expected signs for the true regression parameters. Therefore, it would be extremely helpful if the ridge regression procedure could be modified so as to give small bias for cases involving positive interrelationships and mixed signs of the true regression coefficients. Several possibilities for achieving this objective are explored in the next sections.

## SUPPLEMENTATION OF RIDGE REGRESSION WITH PRIOR INFORMATION

Ridge regression can be considered to be a special case of earlier proposed models, such as the Theil-Goldberger [1961] generalized least squares mixed estimator,

$$(4) \qquad b = (1/\sigma^2 X'X + R'\Psi^{-1}R)^{-1}(1/\sigma^2 X'y + R\Psi^{-1}r).$$

The prior information in (4) is represented by $r = R\beta + v$, where r is a known $g \times 1$ vector ($g \leq p$), R is a known $g \times p$ matrix, and v is a $g \times 1$ vector of errors with $E(v) = 0$, $E(vv') = \Psi$. If $R'\Psi^{-1}R = k/\sigma^2 I_p$ and r is a $p \times 1$ null vector, then (4) reduces to (2) and $b = \hat{\beta}^*$.[1/] (Of course, for the prior information vector, r, equal to a vector of zeroes, it would not seem realistic to assume that r was an unbiased prior estimate of $\beta$, as is assumed for b in (4).)

---

[1/] This fact was first brought to the attention of the author by Don A. Pierce.

Given the above relationship between the ridge estimator and the Theil-Goldberger mixed estimator, it is tempting to try to improve upon the null vector as the prior information vector of $\beta$. In fact, however, the null vector gives surprisingly good results in several situations, as will be shown later. First, however, it should be noted that prior information is often related to the sample information, and this case needs to be handled differently than when the prior vector is independent of the sample data.

## Prior Information Vector Independent of Sample Data

In this case, the researcher may prefer to use the Theil-Goldberger model directly, if the prior vector is considered to be unbiased. However, if the primary motivation is to reduce unduly high variances caused by multicollinearity, then the researcher may want to use ridge regression supplemented by prior information, especially if the researcher is not convinced that his prior vector is unbiased.

As an example, consider the Cobb-Douglas total value production function, $Y = \alpha K^{\beta_1} L^{\beta_2}$, reduced to two inputs, capital and labor, for simplicity. One could estimate

$$b = (X'X + k\lambda)^{-1}(X'y + k\lambda P), \tag{5}$$

where $X'X$ represents the $2 \times 2$ matrix of mean-corrected sums of squares and cross-products, $\lambda$ is a diagonal matrix of order 2 consisting of the sums of squares, and $P$ is a $2 \times 1$ vector of prices for capital and labor.

The proposed prior vector in (5) may not be unbiased since prices expected by managers for capital and labor may not coincide with historical prices used to compute P. Nevertheless, even though in practice the P vector may be in error, a more accurate estimate may still be possible from b, as compared to OLS, if the labor and capital inputs are highly correlated. If the historic input prices, P, are assumed to be known essentially without error, then the variance of b would be computed the same as for the ridge estimator, $\hat{\beta}^*$. However, when the prior information vector is variable or dependent upon the sample data, then the variance of b will usually exceed that of $\hat{\beta}^*$ and should be estimated differently.

## Prior Information Related to Sample Data

In practice, the researcher may not be able to obtain an independent estimate of the prior information vector. In that case, a prior vector related to the sample can be used. One such possibility would be to use the mean of the ordinary least squares (OLS) estimates, $\bar{\hat{\beta}}$, for the elements of the prior vector.[2/] The variance of

$$b = (X'X + k\lambda)^{-1}(X'y + k\lambda\bar{\hat{\beta}}) \tag{5a}$$

will differ from the ridge estimator, $\hat{\beta}^*$. To compute the variance of b, note that the $p \times 1$ mean vector,

$$\bar{\hat{\beta}} = RX'y \tag{6}$$

where R represents a restricted $p \times p$ matrix with p identical rows, where the $i^{\underline{th}}$ element of each row of R is the simple average of all the elements of the $i^{\underline{th}}$ column of $(X'X)^{-1}$. Equation (5a) can be written as

$$\begin{aligned} b &= [(X'X + k\lambda)^{-1}(I_p + k\lambda R)]\, X'y \\ &= AX'y. \end{aligned} \tag{5b}$$

As is well known, the variance of any linear estimator, $b = AX'y$, is equal to $\sigma^2 AX'XA'$ if $E(uu') = \sigma^2 I_n$. (This result follows from the definition of the variance of b and by taking the expected values.) Using these relationships, the variances and mean square error of b and $\hat{\beta}^*$ are presented in Table 1, using the simple two-explanatory-variables experimental models reported by Brown [1973] and summarized by Brown and Beattie [1975].

---

2/ Again, a helpful idea from Don A. Pierce, who suggested this model to the author in 1973, is appreciated. This model also appears to be similar to an estimator analyzed by Benee F. Swindel [1974]. The main objective of Swindel's paper appears to be to prove that there always exists some "good" ridge estimator based on prior information, "good" meaning that the estimator provides a strictly smaller MSE of every non-null linear combination of the slope parameters than does the OLS estimator. Although apparently successful in this regard, Swindel does not attempt to deal with the practical problems involved in actually using and estimating this type of model.

Table 1. Expected Values, Variance, and Mean Square Error for OLS Versus Ridge Regression, and Ridge Regression Supplemented by the Mean Vector, $b = (X'X + k\lambda)^{-1}(X'y + k\lambda\bar{\hat{\beta}})$ where the True Model is $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$; $E(u_i) = 0.5(4) + 0.5(-4)$; and Fixed $X_1$ and $X_2$ take the 5 Values (1,1.2), (2,1.67335), (3,3), (4,4.32665), and (5,4.8)

| | Results for several levels of k | | | | | | |
|---|---|---|---|---|---|---|---|
| | k=0.0 (OLS) | k=0.05 | k=0.10 | k=0.15 | k=0.20 | k=0.25 | k=0.30 |
| $V(\hat{\beta}_1^*)$ | 54.936 | 3.190 | 1.258 | 0.781 | 0.587 | 0.485 | 0.423 |
| $V(b_1)$ | 54.936 | 3.209 | 1.295 | 0.836 | 0.658 | 0.570 | 0.521 |
| $E(\hat{\beta}_1^*)$ | 5.330 | 4.202 | 3.976 | 3.835 | 3.723 | 3.624 | 3.535 |
| $E(\hat{\beta}_2^*)$ | 2.665 | 3.597 | 3.635 | 3.598 | 3.541 | 3.477 | 3.411 |
| $E(b_1)$ | 5.330 | 4.300 | 4.168 | 4.116 | 4.089 | 4.071 | 4.060 |
| $E(b_2)$ | 2.665 | 3.695 | 3.827 | 3.879 | 3.906 | 3.924 | 3.935 |
| Ave. $MSE(\hat{\beta}_1^*)$ | 54.936 | 4.261 | 2.645 | 2.333 | 2.262 | 2.269 | 2.312 |
| Ave. $MSE(b_1)$ | 54.936 | 4.271 | 2.646 | 2.309 | 2.199 | 2.155 | 2.135 |
| $E(\hat{\beta}_1^*)$ | 6.655 | 4.499 | 4.142 | 3.950 | 3.809 | 3.694 | 3.593 |
| $E(\hat{\beta}_2^*)$ | 1.331 | 3.291 | 3.461 | 3.475 | 3.446 | 3.399 | 3.345 |
| $E(b_1)$ | 6.655 | 4.597 | 4.334 | 4.230 | 4.175 | 4.141 | 4.117 |
| $E(b_2)$ | 1.331 | 3.389 | 3.652 | 3.756 | 3.811 | 3.845 | 3.869 |
| Ave. $MSE(\hat{\beta}_1^*)$ | 54.936 | 7.435 | 6.684 | 6.740 | 6.872 | 7.007 | 7.138 |
| Ave. $MSE(b_1)$ | 54.936 | 7.445 | 6.685 | 6.716 | 6.808 | 6.893 | 6.962 |
| $E(\hat{\beta}_1^*)$ | 7.971 | 4.791 | 4.304 | 4.060 | 3.893 | 3.760 | 3.648 |
| $E(\hat{\beta}_2^*)$ | 0.000 | 2.983 | 3.284 | 3.350 | 3.348 | 3.319 | 3.276 |
| $E(b_1)$ | 7.971 | 4.889 | 4.495 | 4.340 | 4.258 | 4.206 | 4.171 |
| $E(b_2)$ | 0.000 | 3.081 | 3.475 | 3.630 | 3.713 | 3.764 | 3.800 |
| Ave. $MSE(\hat{\beta}_1^*)$ | 54.936 | 12.693 | 13.373 | 14.038 | 14.506 | 14.855 | 15.101 |
| Ave. $MSE(b_1)$ | 54.936 | 12.703 | 13.384 | 14.015 | 14.444 | 14.741 | 14.958 |
| $E(\hat{\beta}_1^*)$ | 9.918 | 5.220 | 4.538 | 4.219 | 4.011 | 3.853 | 3.724 |
| $E(\hat{\beta}_2^*)$ | -1.984 | 2.520 | 3.016 | 3.158 | 3.197 | 3.194 | 3.169 |
| $E(b_1)$ | 9.918 | 5.317 | 4.729 | 4.497 | 4.374 | 4.297 | 4.245 |
| $E(b_2)$ | -1.984 | 2.617 | 3.206 | 3.437 | 3.561 | 3.637 | 3.690 |
| Ave. $MSE(\hat{\beta}_1^*)$ | 54.936 | 24.367 | 28.225 | 30.242 | 31.457 | 32.277 | 32.881 |
| Ave. $MSE(b_1)$ | 54.936 | 24.377 | 28.266 | 30.219 | 31.395 | 32.166 | 32.709 |

From the top lines of Table 1, it can be seen that the variances of $\hat{\beta}^*$ and b are very similar, with the variance of $\hat{\beta}^*$ declining somewhat more rapidly as k is increased. However, this small advantage of lower variance for the ridge estimator is offset at the larger values of k by slightly larger biases. For example, for the model $Y_i = \alpha + 5.330X_{1i} + 2.665X_{2i} + u_i$, the expected variance plus expected bias squared, equal to average mean square error (MSE), is slightly lower at k = 0.05 and k = 0.10 for the ridge estimator. Then, for values of k = 0.15 and higher, there is a slightly lower MSE for b, due to its slightly lower bias. (There should be a smaller bias for b at the larger k values since b is being pulled toward the average vector rather than the zero vector. Although b is biased, the sum of the $b_i$ values is unbiased.)

Probably the most striking result of Table 1 is the great similarity of the ridge estimates to the ridge estimates supplemented by the mean vector, $b = \hat{\beta}^* + k(X'X + k\lambda)^{-1}\lambda\hat{\beta}$. These results show that the ridge estimator tends to pull its estimates toward the mean vector, at least in the usually relevant range of k, say $0 < k \leq 0.3$. This fact also fits in with the implication of the theorem of (3). The bias of an estimate pulled toward the mean vector is bound to be larger when the true $\beta_i$ values are more divergent from each other. Thus, higher MSE would be obtained from ridge regression when some of the true $\beta$ values are positive and some negative, as compared to the case where all the signs of the $\beta$ values were the same, assuming the same positive interrelationship among the explanatory variables and about the same $R^2$ for the two models.

At this point, the reader may wonder if he could not do much better than the mean vector, especially for a model like the last one shown in Table 1, $Y_i = \alpha + 9.918X_{1i} - 1.9836X_{2i} + u_i$. Suppose that one was sure that $\beta_1$ was positive and $\beta_2$ negative. As an approximation, one might use $\beta^*$ as the prior vector in (5a), where $\beta^*$ is a restricted OLS estimate such that $\beta_1^* = -\beta_2^*$. If so, $E\beta_1^* \doteq 5.9508$ and $E\beta_2^* = -5.9508$. Intuitively, this vector might be expected to give a lower MSE than the mean vector, $\hat{\beta}$, or lower MSE than the null vector of ridge regression. In fact, however, such is not the case! For example, at k = 0.2, the supplemented ridge estimate,

$$b = (X'X + k\lambda)^{-1}(X'y + k\lambda\beta^*), \tag{7}$$

where $E\beta_1^* = E(-\beta_2^*) = 5.9508$ for the model $Y_i = \alpha + 9.918X_{1i} - 1.9836X_{2i} + u_i$ of Table 1, had MSE(b) $\doteq$ 109.99, as compared to total MSE($\hat{\beta}^*$) = 2(31.457) $\doteq$ 62.91, bottom of Table 1.

It could be argued that $\beta_1^* = -\beta_2^*$ was a poor choice of restriction for $\beta^*$. However, even if one chooses a restriction somewhat closer to the true $\beta$ values, say $\beta_1^* = -2\beta_2^*$, the MSE is still not very impressive. For this case, $E\beta_1^* \doteq 15.3740$, $E\beta_2^* \doteq -7.6870$ (using the same experimental model with $\beta_1 = 9.918$ and $\beta_2 = -1.9836$), and MSE(b) $\doteq$ 62.12 at k = 0.2, almost the same as for the ordinary ridge estimator, MSE($\hat{\beta}^*$) = 2(31.457) $\doteq$ 62.91, Table 1.

While these results are very limited, they do indicate that it is not so easy as it might appear to improve upon ridge regression by supplementing the ridge estimator with a non-zero prior vector. In the preceding model with different signs for $\beta_1$ and $\beta_2$, the reduced bias from using the restricted OLS estimate as the prior vector was offset by the increased variance which resulted. Certainly if a restricted OLS estimate is used as a prior vector, the researcher should compute the new variance, as indicated after Equation (5b), since variances of supplemented ridge estimates can be much larger than for the ordinary ridge estimator, as just illustrated by supplementing the ridge estimator for the model, $Y_i = \alpha + 9.918X_{1i} - 1.9836X_{2i} + u_i$.

Given the problems resulting from attempting to supplement the ridge estimator with restricted OLS prior vectors, a different approach was sought. Inasmuch as the main difficulty with ridge regression arises (under the condition of positive interrelationship among the explanatory variables) when there are unlike signs of the true regression coefficients, the idea of obtaining ridge estimates of those coefficients with the same expected sign seemed highly desirable. A "partitioned" ridge regression procedure for accomplishing this objective is presented next.

## "PARTITIONED" RIDGE REGRESSION

For sake of exposition, assume that the model to be fitted has p mean-corrected explanatory variables that are all positively interrelated and that the true,

known sign of q coefficients of one block of variables differs from the true, known sign of the (p-q) remaining variables. Then, for lack of a better term, the proposed "partitioned" ridge regression procedure consists of the following steps:

1. Fit the full model, $Y = \bar{X}_1\beta_1 + \bar{X}_2\beta_2 + u$ by OLS, where the q variables with one expected sign are contained in a block of variables denoted by $\bar{X}_1$, and $\bar{X}_2$ contains the remaining (p-q) block of variables with true coefficients of opposite sign.

2. Construct a new variable, $Z_1 = Y - \bar{X}_2\hat{\beta}_2$, then fit $Z_1 = \bar{X}_1\beta_1$ by regular ridge regression. Similarly, define $Z_2 = Y - \bar{X}_1\hat{\beta}_1$ and fit $Z_2 = \bar{X}_2\beta_2$ by regular ridge regression.

3. Combine these results to obtain $Y^* = \bar{X}_1 b_1^* + \bar{X}_2 b_2^*$, from

(8) $$b_1^* = (\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1} X_1'Z_1 \quad \text{and} \quad b_2^* = (\bar{X}_2'\bar{X}_2 + k\lambda_2)^{-1} X_2'Z_2.$$

In (8), $\bar{X}'\bar{X}$ refers to the mean-corrected sums of squares and cross-products for the block of q explanatory variables with true coefficients of like sign, as defined in Step 1 above, and similarly for $\bar{X}_2'\bar{X}_2$. The $q \times q$ diagonal matrix $\lambda_1$ consists of the main diagonal elements of $\bar{X}_1'\bar{X}_1$, and the $(p-q) \times (p-q)$ matrix $\lambda_2$ consists of the main diagonal elements of $\bar{X}_2'\bar{X}_2$.

It should be noted that if one of the matrices, $\bar{X}_1'\bar{X}_1$ or $\bar{X}_2'\bar{X}_2$, is orthogonal, or nearly so, partitioned ridge regression can not greatly reduce MSE as compared to the OLS estimates for the variables of the orthogonal matrix. This point will be elaborated upon with respect to the numerical examples presented later.

It appears intuitively that the variance of $b_1^*$ in (8) will be relatively lower if the OLS estimate of $\beta_2$ in Step 1 is relatively precise, and similarly for $b_2^*$. This result also follows from the derivation of the variance of $b_1^*$ and $b_2^*$. As defined in (8), the non-standardized $p \times 1$ vector, $b_1^*$, is:

$$(9) \qquad b_1^* = (\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}\,\bar{X}_1'Z_1 = (\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}\,\bar{X}_1'\,(Y - \bar{X}_2\hat{\beta}_2)$$

$$= (\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}\,\bar{X}_1'\,(\bar{X}_1\beta_1 + \bar{X}_2\beta_2 - \bar{X}_2\hat{\beta}_2 + u)$$

$$= (\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}\,\bar{X}_1'\,[\bar{X}_1\beta_1 - \bar{X}_2(\hat{\beta}_2 - \beta_2) + u].$$

The $n \times 1$ error term, $u$, is assumed to have expected value equal to zero, and $E(u'u) = \sigma^2 I_n$. Assuming $\bar{X}_1$ and $\bar{X}_2$ are fixed,

$$(10) \qquad E(b_1^*) = (\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}\,[(\bar{X}_1'\bar{X}_1 + k\lambda_1)\,\beta_1 - k\lambda_1\beta_1]$$

$$= \beta_1 - k(\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}\,\lambda_1\beta_1.$$

Thus, the estimate of $b_1^*$ will be biased only to the extent caused by the elements of $\beta_1$, and this bias will be relatively small since, by hypothesis, all elements of $\beta_1$ are of the same sign, and all variables in block $X_1$ are positively related. Thus, as discussed earlier with (3), relatively small bias can be expected. By definition, the variance-covariance matrix for $b_1^*$ is:

$$(11) \quad \mathrm{Var}(b_1^*) = E(b_1^* - Eb_1^*)(b^* - Eb_1^*)'$$

$$= E\{(\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}\,\bar{X}_1'[u - \bar{X}_2(\hat{\beta}_2 - \beta_2)]\}$$

$$\times\ \{(\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}\,\bar{X}_1'[u - \bar{X}_2(\hat{\beta}_2 - \beta_2)]\}'$$

$$= (\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}\,\bar{X}_1'(Euu')\,\bar{X}_1(\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}$$

$$+ (\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}\,\bar{X}_1'\bar{X}_2[E(\hat{\beta}_2 - \beta_2)(\hat{\beta}_2 - \beta_2)']\,\bar{X}_2'\bar{X}_1\,(\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}.$$

Since $E(uu') = \sigma^2 I_n$, by assumption,

$$(12) \quad \mathrm{Var}(b_1^*) = \sigma^2(\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}\,\bar{X}_1'\bar{X}_1\,(\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}$$

$$+ (\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}\,\bar{X}_1'\bar{X}_2[\mathrm{Var}(\hat{\beta}_2)]\,\bar{X}_2'\bar{X}_1\,(\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}.$$

The first term of (12) is simply the variance of the regular ridge estimate of $\hat{\beta}_1^* = (\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}\bar{X}_1'Z_1$. The last term of (12) is the variance added to $Var(\hat{\beta}_1^*)$ by constructing $Z_1 = Y - \bar{X}_2\hat{\beta}_2$ before making the ridge estimation, as defined in Step 2. It is clear that the last term in (12) vanishes if all variables in $\bar{X}_2$ are uncorrelated with the variables of $\bar{X}_1$. However, even though the $\bar{X}_2$ variables will usually be correlated with $\bar{X}_1$ variables, the last term of (12) will be relatively small if the variance of $\hat{\beta}_2$ is small. Of course, a substantially smaller $MSE(b_1^*)$ relative to MSE of the regular ridge estimate of $\beta_1$ will result only if the last term of (12) is substantially smaller than the bias incurred from the full ridge regression.

The preceding material will next be illustrated with a simple experimental model.

Results From A Simple Monte Carlo Experiment

The experimental equation was:

$$(13) \qquad Y_i = 18 + 3X_{1i} + 2X_{2i} - 6X_{3i} + u_i$$

where $X_1$, $X_2$, and $X_3$ always took the fixed values, (0,1,0), (1,0,2), (2,2,1), (8,8,9), (9,10,8), and (10,9,10). Since $E(u_i) = 0.5(-4) + 0.5(4)$ for the experiments summarized in Table 2, there are $2^6 = 64$ combinations of the binomial error term, which results in 64 possible outcomes or samples.[3/] However, some of the samples give the same parameter estimates, resulting in only 27 unique sets of estimates, as shown in Table 2. (Some of the samples within unique sets do give different estimates of $\sigma^2$, however.)

---

3/ The outcomes were generated as follows. First, note that if $\alpha = 18$, the six expected values of $Y_i$ were 20, 9, 22, 4, 17, and 6, respectively. The error terms, $u_i$, of the first experiment were all negative, or $u_1 = -4$, $u_2 = -4$, $\cdots$, $u_6 = -4$, giving values of $Y_i$ of 16, 5, 18, 0, 13, and 2, respectively. For the second trial, $u_1 = u_2 = u_3 = u_4 = u_5 = -4$, and $u_6 = +4$, giving $Y_i$ values of 16, 5, 18, 0, 13, and 10, respectively. The third trial had $u_1 = u_2 = u_3 = u_4 = u_6 = -4$, and $u_5 = +4$, giving $Y_i$ values of 16, 5, 18, 0, 21, and 2, respectively. This truth table type of process was repeated until all 64 possible outcomes had been generated.

Table 2. Distribution of Estimated β Values for OLS Versus "Partitioned" Ridge Regression, where $Y_i = \alpha + 3X_{1i} + 2X_{2i} - 6X_{3i} + u_i$; $E(u_i) = 0.5(4) + 0.5(-4)$; and Fixed $X_1$, $X_2$, and $X_3$ Take the Six Values (0,1,0), (1,0,2), (2,2,1), (8,8,9), (9,10,8), and (10,9,10)

| Relative frequency | k = 0.0 (OLS) | | | k = 0.2 (Partitioned Ridge Estimates) | |
|---|---|---|---|---|---|
| | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $b_1^*$ | $b_2^*$ |
| 1/64 | 10.667 | -0.333 | -11 | 5.1927 | 4.1927 |
| 2/64 | 11.000 | -2.000 | -10 | 4.6781 | 3.4962 |
| 2/64 | 7.000 | 2.000 | -10 | 4.3144 | 3.8599 |
| 1/64 | 11.333 | -3.667 | - 9 | 4.1635 | 2.7998 |
| 4/64 | 7.333 | 0.333 | - 9 | 3.7998 | 3.1635 |
| 1/64 | 3.333 | 4.333 | - 9 | 3.4362 | 3.5271 |
| 2/64 | 7.667 | -1.333 | - 8 | 3.2852 | 2.4671 |
| 2/64 | 6.333 | 1.333 | - 8 | 3.7089 | 3.2544 |
| 2/64 | 3.667 | 2.667 | - 8 | 2.9216 | 2.8307 |
| 4/64 | 6.667 | -0.333 | - 7 | 3.1943 | 2.5580 |
| 1/64 | 4.000 | 1.000 | - 7 | 2.4070 | 2.1343 |
| 4/64 | 2.667 | 3.667 | - 7 | 2.8307 | 2.9216 |
| 2/64 | 7.000 | -2.000 | - 6 | 2.6797 | 1.8616 |
| 8/64 | 3.000 | 2.000 | - 6 | 2.3161 | 2.2252 |
| 2/64 | -1.000 | 6.000 | - 6 | 1.9525 | 2.5888 |
| 4/64 | 3.333 | 0.333 | - 5 | 1.8015 | 1.5288 |
| 1/64 | 2.000 | 3.000 | - 5 | 2.2252 | 2.3161 |
| 4/64 | -0.667 | 4.333 | - 5 | 1.4379 | 1.8924 |
| 2/64 | 2.333 | 1.333 | - 4 | 1.7106 | 1.6197 |
| 2/64 | -0.333 | 2.667 | - 4 | 0.9233 | 1.1960 |
| 2/64 | -1.667 | 5.333 | - 4 | 1.3470 | 1.9833 |
| 1/64 | 2.667 | -0.333 | - 3 | 1.1960 | 0.9233 |
| 4/64 | -1.333 | 3.667 | - 3 | 0.8324 | 1.2869 |
| 1/64 | -5.333 | 7.667 | - 3 | 0.4687 | 1.6505 |
| 2/64 | -1.000 | 2.000 | - 2 | 0.3178 | 0.5905 |
| 2/64 | -5.000 | 6.000 | - 2 | -0.0459 | 0.9541 |
| 1/64 | -4.667 | 4.333 | - 1 | -0.5605 | 0.2577 |
| Average value | 3.000 | 2.000 | -6.000 | 2.3161 | 2.2252 |
| Variance | 16.167 | 5.500 | 5.500 | 1.6188 | 0.7380 |
| Bias squared | 0.000 | 0.000 | 0.000 | 0.4677 | 0.0507 |
| M.S. error | 16.167 | 5.500 | 5.500 | 2.0865 | 0.7887 |

An advantage of the experimental scheme used in Table 2 is that all possible outcomes are obtained, so that the averages over the entire experiment correspond exactly (except for slight rounding error) to the mathematical expectations of the parameters, their variances, and their biases. A possible disadvantage is that the relatively small sample size does not permit as wide a range of extreme values, even though the expected values of the variances and biases are the same as for any other type of disturbance term with $\sigma^2 = 16$. The dispersion of the OLS estimates of $\beta_3$ is shown in Figure 1.

Before discussing the partitioned versus the regular or full-model ridge regression estimates, the behavior of the OLS estimates of the parameters of (13) in Table 2 should be noted. The OLS estimates of $\beta_1$ and $\beta_2$ were highly unstable, with $\hat{\beta}_1$ taking the wrong (negative) sign 20/64 = 5/16 of the time and $\hat{\beta}_2$ taking the wrong sign 13/64 of the time! The main reason for this instability was high intercorrelation among the explanatory variables. For example, correlation between $X_1$ and $X_2$, $r_{12}$, was $r_{12} = 0.98$. Similarly, $r_{13} = 0.98$, and $r_{23} = 0.94$. These high intercorrelations resulted in large main diagonal elements of the inverted correlation matrix, the so-called "Variance Inflation Factors" (VIF). The VIF for $X_1$, $X_2$, and $X_3$ were 101.042, 34.375, and 34.375, respectively.

These high variances of the OLS estimates can, of course, be reduced by the use of regular ridge regression on the full model of (13). However, a rather serious bias quickly results, as the following figures show:

| Value of k | $E(\hat{\beta}_1^*)$ | $E(\hat{\beta}_2^*)$ | $E(\hat{\beta}_3^*)$ | $V(\hat{\beta}^*)$ | Total MSE $(\hat{\beta}^*)$ |
|---|---|---|---|---|---|
| 0.0 (OLS) | 3.000 | 2.000 | -6.000 | 27.167 | 27.167 |
| 0.05 | 0.069 | 1.676 | -2.688 | 1.174 | 20.840 |
| 0.10 | -0.106 | 1.090 | -1.910 | 0.518 | 27.208 |
| 0.15 | -0.167 | 0.771 | -1.515 | 0.310 | 31.966 |
| 0.20 | -0.320 | 0.573 | -1.273 | 0.214 | 34.591 |

Although lowest MSE fell at k = 0.01 (to nearest one-hundredth), giving $MSE(\hat{\beta}^*) = 12.59$, if one selected k = 0.10 or k = 0.20, a higher MSE would result than for OLS! Thus, the model is somewhat unsuitable for estimation by

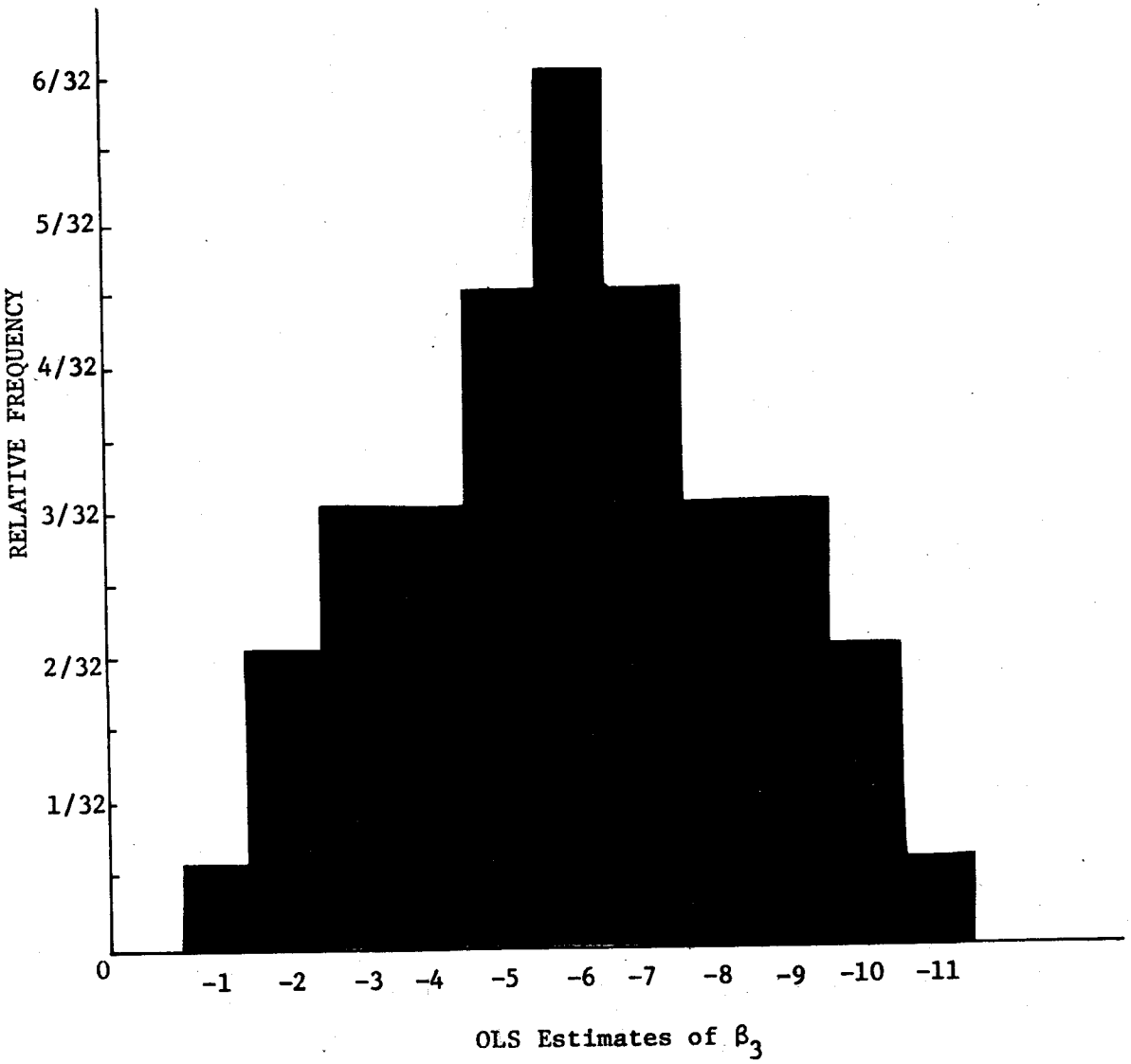



Figure 1. Distribution of $\beta_3$ across all samples for the model, $Y_i = 18 + 3X_{1i} + 2X_{2i} - 6X_{3i} + u_i$, $E(u_i) = .5(4) + .5(-4)$, $n = 6$, where fixed $X_1$, $X_2$, and $X_3$ take the values (0,1,0), (1,0,2), (2,2,1), (8,8,9), (9,10,8), and (10,9,10).

either OLS or regular ridge regression. However, use of "partitioned" ridge regression gives a surprisingly good result, as indicated by the estimates of the last two columns of Table 2. In contrast to OLS $\hat{\beta}_1$, which took the wrong (negative) sign 20/64 of the time, the partitioned ridge estimator $b_1^*$ at k = 0.2 took the wrong sign only 3/64 (less than 5 percent) of the time. Similarly, whereas $\hat{\beta}_2$ had the wrong sign 13/64 $\doteq$ 20.3 percent of the time, the partitioned ridge estimate $b_2^*$ had the correct sign 100 percent of the time.

The expected values, variances, and MSE of $b_1^*$ and $b_2^*$ for various k values were as follows:

| Value of k | $E(b_1^*)$ | $E(b_2^*)$ | $V(b_1^*)$ + $V(b_2^*)$ | $MSE(b_1^*)$ + $MSE(b_2^*)$ |
|---|---|---|---|---|
| 0.0 (OLS) | 3.000 | 2.000 | 21.667 | 21.667 |
| 0.10 | 2.463 | 2.296 | 2.944 | 3.320 |
| 0.20 | 2.316 | 2.225 | 2.357 | 2.875 |
| 0.30 | 2.202 | 2.140 | 2.085 | 2.741 |
| 0.40 | 2.104 | 2.056 | 1.889 | 2.695 |
| 0.50 | 2.015 | 1.977 | 1.728 | 2.698 |

In computing the variances of $b_1^*$ and $b_2^*$, care must be taken not to forget the second term of (12), which must be added to the usual variance of the two-explanatory-variable ridge regression model. Because of this added term, the variance of the partitioned ridge estimator is a relatively larger component of MSE, especially at the larger k values. Consequently, $MSE(b^*)$ continued to decline over a larger range of k than for similar two-explanatory-variable models fitted by regular ridge regression, reported by Brown [1973]. Although more research is needed, it appears that larger k values can be used to obtain lower MSE for partitioned ridge regression than for regular ridge regression.

Although MSE of the estimates of the model in Table 2 are much lower for partitioned ridge regression than for OLS or regular ridge regression, several legitimate questions are in order at this point.

(1) What would be the effect on the outcome if all the true coefficients were proportionately larger or smaller, given the same X and u values?

(2) How would the outcome be affected if the intercorrelations among the explanatory variables were lower or higher?

(3) How does partitioned ridge regression compare with variable deletion, in coping with the multicollinearity of the model of (13) and Table 2?

Question 1: If the true $\beta$ values of (13) and Table 2 were all proportionately larger, OLS would do somewhat better relative to regular ridge regression since the bias increases directly with the absolute magnitude of the true $\beta$ values, as shown by (3). Similarly, the bias of the partitioned ridge estimates would also increase, but should always remain much smaller than for the regular ridge estimates.

Question 2: The relative advantage of both partitioned and regular ridge regression increases (decreases) as the degree of multicollinearity increases (decreases), since the variance can always be effectively reduced with ridge regression. On the other hand, the bias depends more on the true $\beta$ coefficients, as can be seen from (3), and discussed by Brown and Beattie [1975, pp. 23-26].

Question 3: Using variable deletion on the model of Table 2, the variable $X_1$ would be most often deleted since the ratio of $\hat{\beta}_1$ to its expected standard error is lowest most often in Table 2. Deleting $X_1$, then the total variance for the OLS estimate of $\beta_2$ and $\beta_3$ will be reduced to $16(0.08591065)(2) \doteq 2.7491$. However, the expected values of the OLS estimates of $\beta_1$, $\beta_2$, and $\beta_3$ will now be biased, being 0, 3.515464, and -4.484536, respectively, and the bias squared would be $(0-3)^2 + (3.515464 - 2)^2 + (-4.484536 + 6)^2 \doteq 13.5933$. Thus, the total MSE would be $2.7491 + 13.5933 \doteq 16.3424$. Thus, MSE for the OLS estimate of the model of Table 2, with $\beta_1$ set equal to zero, is less than the variance of the full OLS estimate, which was $16.167 + 5.500 + 5.50 \doteq 27.167$. However, the MSE of the deleted variable model is considerably higher than for the partitioned ridge regression model, where MSE at $k = 0.2 \doteq 2.0865 + 0.7887 + 5.5000 \doteq 8.3752$.

It should be noted that MSE of the partitioned ridge estimate is slightly lower at higher values of k. Also, the MSE could be reduced slightly by fitting a partitioned ridge estimate of $\beta_3$, using the same general equation of (9). However, not much reduction in variance would be expected since, for the model of Table 2, the (p-q) × (p-q) matrix $\overline{X}_2'\overline{X}_2$ consists of a single number, namely, $X_3'X_3$. Nevertheless, fitting $b_2^*$, as defined in (8), a minimum $MSE(b_2^*)$ is obtained for $k \doteq 0.15$, where $V(b_2^*) \doteq 4.1588$, $bias^2(b_2^*) \doteq (-5.21739 + 6)^2 \doteq 0.6125$, and $MSE(b_2^*) \doteq 4.7713$. Thus, a reduction of MSE of (5.5 - 4.7713) = 0.7287 was obtained, only about 13 percent. However, had $\overline{X}_2$ contained more than one variable, $X_3$, a bigger reduction in MSE would be obtained, assuming that the variables were not orthogonal. This fact will be illustrated in the next section, where partitioned ridge regression is applied to a simple four-explanatory-variable model.

## Results from a Four-Explanatory-Variable Model

For simplicity, consider a four-explanatory-variable model where there is equal correlation between each variable, say $r_{12} = r_{13} = r_{14} = r_{23} = r_{24} = r_{34} = 0.95$. Furthermore, suppose there are 25 observations and the overall $R^2$ of regression is 0.49875, yielding an overall regression $F_{20}^4 = 4.975$ that is significant at P = 0.01. If we set $\beta_1 = \beta_2 = -\beta_3 = -\beta_4$ and solve, we obtain standardized $\beta_1 = \beta_2 = 1.57916117$ and $\beta_3 = \beta_4 = -1.57916117$.

Using OLS on the above model, the Variance Inflation Factor (VIF) is 15.0649 for each explanatory variable, a moderately high level of multicollinearity corresponding to $r_{19} \doteq 0.96624$ in a two-variable-explanatory model. Multiplying $\sigma^2 = (1-R^2)/20$ times the VIF gives $V(\hat{\beta}_i) \doteq 0.37765$, and $t \doteq 2.57$. $MSE(\hat{\beta}) \doteq 1.51$.

If regular ridge regression is attempted on the above model, very poor results are obtained. For k = 0.1, a badly biased ridge coefficient is estimated, $\hat{\beta}_1^* = \hat{\beta}_2^* \doteq 0.5264 = -\hat{\beta}_3^* = -\hat{\beta}_4^*$. For k = 0.1, $MSE(\hat{\beta}^*) \doteq 4.6066$, over three times that of OLS!

Admittedly, a smaller $MSE(\hat{\beta}^*)$ can be obtained by using a smaller value of k, although in practice one would not know just what value of k to select. However, for this model the optimal value of k to the nearest thousandth is k = 0.008, giving $MSE(\hat{\beta}^*) \doteq 1.3138$. But even at this optimal value of k the reduction in

MSE is not great, being about 1.3138 ÷ 1.51026 ≐ 87 percent that of OLS.

In contrast to the poor results with regular ridge regression, very good results can be achieved with partitioned ridge regression over a wide range of k values. The partitioned ridge estimates, their variances, and MSE were as follows:

| Value of k | Partitioned ridge estimate of $\beta_1 = \beta_2$ | Partitioned ridge estimate of $\beta_3 = \beta_4$ | Individual variances | Individual MSE | Total MSE($b^*$) |
|---|---|---|---|---|---|
| 0.0 (OLS) | 1.5791 | -1.5791 | 0.37756 | 0.37756 | 1.51026 |
| 0.1 | 1.5021 | -1.5021 | 0.14270 | 0.14270 | 0.59456 |
| 0.2 | 1.4323 | -1.4323 | 0.11445 | 0.13603 | 0.54410 |
| 0.4 | 1.3104 | -1.3104 | 0.09050 | 0.16275 | 0.65099 |

For k = 0.2, total MSE of the partitioned ridge estimator was 0.5441, only about 36 percent that for OLS. Even at k = 0.4, MSE($b^*$) was only about 43 percent that of OLS, still a substantial reduction. Thus, partitioned ridge regression gives a much better result than OLS or regular ridge regression over a wide range of k values.

It should be noted that the preceding model is favorable for partitioned ridge regression in that the β values within blocks are of the same standardized magnitude. A larger bias would result at a given value of k if the true β values within blocks were of unequal magnitude, although not so large as if the β values within a block were of unlike sign. It should also be pointed out that the preceding model would yield even better results for partitioned ridge regression if the correlation between variables of different blocks were lower than the correlation between variables within blocks; that is, if $r_{13}$, $r_{14}$, $r_{23}$, and $r_{24}$ were less than $r_{12} = r_{34} = 0.95$. Of course, the converse is also true. A higher variance from the second term of Equation (12) would result from higher values of $r_{13}$, $r_{14}$, $r_{23}$, and $r_{24}$. In fact, the unfavorable situation for partitioned ridge regression of higher correlation between variables of different blocks is illustrated by the production function example presented next. However, before applying partitioned and regular ridge regression to an empirical situation, some of

the difficulties of interpretation and limitations of previous applications of ridge regression to empirical problems should be examined in order to avoid some of the limitations of these earlier applications.

## Difficulties of Interpretation when Ridge Regression is Applied to Empirical Problems

Although of value in illustrating the use of ridge regression, Hoerl and Kennard [1970b] do not attempt to actually assess the MSE of their ridge estimates in order to compare MSE($\hat{\beta}^*$) with OLS. In fact, such a comparison probably is not possible, since reliable estimates of the true $\beta$ parameters apparently were not available. A similar criticism can be levelled at the application of ridge regression by McDonald and Schwing [1973]. Thus, surprisingly little can be definitely concluded from these empirical examples. Further analyses by Mallows [1973] and Farebrother [1975] question the value of k used by Hoerl and Kennard [1970b], suggesting that much smaller values of k might have yielded ridge estimates with lower MSE.

Given the limitations of the preceding applications, how can applications be made more meaningful in comparing the expected MSE associated with various estimators?

The first requirement for making better inferences from empirical problems is, surprisingly, that one should first start with experimental or survey data that will provide accurate parameter estimates with OLS. One needs to start with very good data to be able to measure bias squared. But from these data the second requirement is to purposely select a subset of data which provides high intercorrelation among the explanatory variables. Thus, one can simulate conditions conducive to the use of biased linear estimation.

Interestingly, fulfilling the two preceding requirements is still not quite enough, because analysis of a given subset of data represents only one sample. Variation of the observed dependent variable for the selected subset will usually give highly variable and unstable results.

What should be done? One way of coping with the difficulty would be to add an error term, $u_i$, to the OLS predicted values for the subset (using the model

fitted from all the data), then to take many samples where $u_i$ could be normally and independently distributed with mean zero and known variance. (One advantage of such a procedure would be that various methods for selecting k could be evaluated.) However, such a procedure would also be fairly expensive and time-consuming. Fortunately, the expected variances and biases that would occur across all possible samples can be approximated without doing the actual experiments, using the following argument: Assuming that the hypothesized very good OLS estimates are available from the data of the entire experiment, it is reasonable to use the OLS estimates as proxies for the true parameters. Then, for given values of k, it is possible to compute the expected parameter estimates and variances, using the usual estimating equations and by using $\hat{\sigma}^2$ from the OLS fit of the total experiment as a proxy for $\sigma^2$.

Thus, the three requirements or steps for making valid inferences from an empirical problem are the following:

1. Start with a good design or data source that provides reliable OLS estimates of the true parameters.

2. Select a subset such that the explanatory variables become more highly intercorrelated, similar to situations that are expected in practice.

3. Use the reliable OLS estimates as proxies for the true parameters, then compute the expected MSE that would result from all possible samples for the various estimators being tested.

The above steps will next be illustrated with the analysis of a production (yield-fertilizer response) function.

## Application of Regular Ridge Regression to An Empirical Production Function

Basic data for this analysis were taken from Table A-14 [Heady, et al., 1955, p. 330]. However, since the original experiment [Heady, et al., 1955] was nearly orthogonal in terms of the phosphorus (P) and nitrogen (N) treatments, a subset of 21 treatments out of the total 57 treatments was selected for analysis.

The 21 treatments were purposely selected along or near the main diagonal of Table A-14 in order to cause a high positive correlation between the P and N treatments, $r \doteq 0.911$. Thus, nonorthogonal data, similar to most economic data, were obtained.[4/] (The selected treatments and yields are given in Appendix Table 1.)

One advantage, as will be seen later, in analyzing a selected subset from this particular experiment was that the OLS estimates fitted to the entire experiment were unusually precise. The quadratic form of the production function, fitted by OLS to all 57 treatments and 114 observations, was:

$$(14) \qquad \hat{Y} = -7.510 + \underset{(.0635)}{0.6638P} + \underset{(.0635)}{0.5843N} - \underset{(.000176)}{0.001797P^2}$$

$$- \underset{(.000176)}{0.001581N^2} + \underset{(.000155)}{0.0008113PN}.$$

$$R^2 = 0.832$$

In (14), $\hat{Y}$ refers to the yield of corn in bushels per acre, P denotes pounds of $P_2O_5$ applied per acre, and N refers to pounds of elemental N applied per acre. (For more details, cf. Heady, et al., 1955.) Standard errors are given in parentheses below the corresponding regression coefficients. Values of t ranged from 5.24 to 10.46, indicating very accurate parameter estimates.

In contrast to the reliable estimates of (14), much worse results were obtained when the same model of (14) was fitted by OLS to the nonorthogonal subset of 21 treatments and 42 observations (given in Appendix Table A-1):

$$(15) \qquad \hat{Y} = 16.79 + \underset{(.1333)}{0.6678P} + \underset{(.1347)}{0.3419N} - \underset{(.00104)}{0.001423P^2}$$

$$- \underset{(.000940)}{0.0008198N^2} + \underset{(.00186)}{0.00009386PN}.$$

$$R^2 = 0.893$$

---

[4/] Little reduction in MSE would be expected from using either regular or partitioned ridge regression on the nearly orthogonal data of the complete experiment. Also, very little knowledge would be gained about the value of ridge regression for highly multicollinear situations.

Values of t for the $N^2$ and interaction term, PN, were quite low, especially for PN with $t \doteq 0.05$. Main reason for the unreliability of estimation in (15) was multicollinearity, brought about by purposely selecting a subset of observations with high correlation between the two inputs, P and N. The so-called VIF (the main diagonal elements of the inverted correlation matrix) were the following:

| Explanatory variable | VIF for (14) | VIF for (15) |
|---|---|---|
| P | 14.6 | 41.0 |
| N | 14.6 | 44.4 |
| $P^2$ | 12.6 | 295.0 |
| $N^2$ | 12.6 | 253.1 |
| PN | 5.5 | 892.1 |

As would be expected, a serious problem of multicollinearity was created (purposely) by selecting the nonorthogonal subset, especially for the squared and interaction terms.

The natural inclination at this point might be to apply partitioned and regular ridge regression to the values of the explanatory variables and $Y_i$ values of the subset used in fitting (15). But that would be of very limited interest, as discussed in the preceding section, since the $Y_i$ values of the subset represent only <u>one</u> sample. We do know that if we took all possible samples of $Y_i$ values corresponding to the subset of explanatory variables in (15), we would obtain an average estimate fairly close to the reliable OLS estimates of (14). Therefore, it is reasonable to make the analysis, assuming that the estimates of (14) are the expected values of the parameters, without being very far off the mark. Using this assumption, and the relationship

$$\hat{\beta}'(X'X)\hat{\beta} = R^2 \sum y_i^2, \tag{16}$$

we obtain an expected sum of squares due to regression across all possible samples of about 90,349.32, using $\hat{\beta}$ from (14) and with X'X being the mean-corrected sums of squares and cross-products of the subset explanatory variables of (15). Similarly, assuming $\hat{\sigma}^2 \doteq 377.1199$, from (14), is the true $\sigma^2$, the expected

deviations from regression for the subset of 42 across all possible samples would be 36(377.1199) ≐ 13,576.32. Thus, $E(\sum y_i^2)$ ≐ 90,349.32 + 13,576.32 ≐ 103,925.64.

Using the above data, it is relatively easy to compute the expected value of the non-standardized ridge estimates at various levels of k. However, there is some advantage in presenting the variances and MSE in terms of the standardized ridge regression estimates, at least for empirical problems. (If not presented in standardized form, MSE estimates of the coefficients of some variables would receive undue weight in estimating total MSE($\hat{\beta}^*$).[5/]

Due to the high VIF, variances of the OLS estimates (first set of estimates in Table 3) were quite large with a total estimated MSE of 5.5357. Smallest MSE, to the nearest one-thousandth value of k, was obtained at k = 0.002, giving estimated total MSE($\hat{\beta}^*$) ≐ 1.84. However, it should be noted that, thus far, there has been no method demonstrated which will assure such an accurate selection of optimal k when the true parameter values are unknown. However, one promising method is to compute the u statistic [Wallace, 1972] to test the hypothesis that the ridge estimates are lower in weak MSE. However, for k = 0.01, only a low value of u ≐ 0.29 is obtained. Not until k = 0.09 and u = 5.14 would one reject the hypothesis that the ridge estimates are lower in weak MSE, using the tabulated values [Goodnight and Wallace, 1972] for 5 and 36 degrees of freedom at P = 0.05. But at k = 0.09, estimated MSE for $\hat{\beta}^*$ rose to 3.196.

Even at k = 0.002 in Table 3, the ridge estimate of the important interaction term, PN, takes a negative sign, but this implied negative PN interaction is contrary to other P-N fertilizer experiments with corn. The negative ridge coefficient for PN also contradicts the relatively precise and unbiased OLS estimate (fitted to all 114 observations) in (14). The ridge-estimated coefficients for $P^2$ and $N^2$ in Table 2 have also moved in the wrong direction, becoming smaller in absolute magnitude, even for k equal to only 0.002.

---

[5/] Weighting was not a problem with the regular ridge estimates of (13) and Table 2 because each explanatory variable was constructed so as to have a mean-corrected sum of squares equal to 100.

Table 3. Standardized Ridge Regression Estimates Corresponding to Expected Values of Non-standardized Estimates of Iowa Corn Production Function Coefficients, Along with Estimated Variance and MSE for Various k Values

| Type of estimate | Explanatory variable | | | | | |
|---|---|---|---|---|---|---|
| | P | N | $P^2$ | $N^2$ | PN | Sum[b] |
| Coefficient @ k = 0.0... | 1.32955 | 1.20568 | -1.24313 | -1.11630 | 0.54296 | -- |
| Variance @ k = 0.0..... | 0.1488 | 0.1610 | 1.0703 | 0.9185 | 3.2371 | 5.5357 |
| Bias squared @ k = 0.0[a] | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Estimated MSE..... | 0.1488 | 0.1610 | 1.0703 | 0.9185 | 3.2371 | 5.5357 |
| Coefficient @ k = 0.002. | 1.2644 | 1.2030 | -0.7531 | -0.6677 | -0.2976 | -- |
| Variance @ k = .002..... | 0.0983 | 0.1083 | 0.1245 | 0.1284 | 0.2324 | 0.6919 |
| Bias squared @ k = .002[a] ..... | 0.0043 | 0.0000 | 0.2401 | 0.2012 | 0.7050 | 1.1506 |
| Estimated MSE..... | 0.1026 | 0.1083 | 0.3646 | 0.3296 | 0.9374 | 1.8424 |
| Coefficient @ k = 0.03.. | 0.9816 | 0.9487 | -0.4286 | -0.3562 | -0.3933 | -- |
| Variance @ k = 0.03..... | 0.0162 | 0.0159 | 0.0110 | 0.0118 | 0.0054 | 0.0603 |
| Bias squared @ k = 0.03[a] ..... | 0.1210 | 0.0660 | 0.6634 | 0.5778 | 0.8765 | 2.3047 |
| Estimated MSE..... | 0.1373 | 0.0819 | 0.6744 | 0.5896 | 0.8819 | 2.3651 |
| Coefficient @ k = .10... | 0.6590 | 0.6465 | -0.2138 | -0.1566 | -0.1970 | -- |
| Variance @ k = .10..... | 0.0053 | 0.0050 | 0.0043 | 0.0042 | 0.0017 | 0.0204 |
| Bias squared @ k = 0.10[a] ..... | 0.4496 | 0.3127 | 1.0595 | 0.9210 | 0.5475 | 3.2903 |
| Estimated MSE..... | 0.4549 | 0.3177 | 1.0637 | 0.9252 | 0.5492 | 3.3107 |

[a] Estimates of bias squared are based upon the assumption that the coefficients in (14), estimated from all 114 observations, are the true parameters.

[b] Sums may not check exactly, due to rounding.

Another possibility for selecting k would be to stop at the first value of k for which Wallace's [1972] u statistic has P ≐ 0.50, following Burt [1974]. The first value of u not to exceed P = 0.50 (to the nearest one-hundredth value of k) was at k = 0.03, the third set of estimates in Table 3. Using this procedure, an estimated MSE of 2.3651 is obtained.[6/]

Higher values of k give progressively higher estimated MSE. For example, at k = 0.10 in the lower part of Table 3, a total estimated MSE ≐ 3.31 was computed.

The reason for the excessive bias of the ridge estimate of the PN variable coefficient can be seen from earlier Equation (3). If PN is fitted as an auxillary ridge function of the four remaining explanatory variables, say at k = 0.03, the following equivalent of (3) is obtained:

$$(17) \qquad E(\hat{\beta}_5^* - \beta_5) = 0.03(22.2025)\,[0.058344(1.32955) + 0.084867(1.20568) + 0.450433(-1.24313) + 0.432285(-1.11630) - (0.54296)] \doteq -0.93622.$$

This estimated bias checks with the assumed true $\beta_5$ value, given in the first line of Table 4, and the ridge estimate of $\beta_5$ for k = 0.03 in Table 4, since -0.39326 - 0.54296 = -0.93622. The main problem in (17) is that the PN variable is a strong positive ridge function of the $P^2$ and $N^2$ explanatory variables, and the expected true β values of $P^2$ and $N^2$ are negative, being -1.24313 and -1.11630,

---

[6/] Actually, the correct interpretation is that if one always selected k = 0.03, the non-standardized ridge estimate, $\hat{\beta}_j^*$, would have expected values corresponding to $(\sqrt{E(\sum y^2)} \div \sqrt{\sum x_j^2}) \doteq (322.375 \div \sqrt{\sum x_j^2})$ times the $j^{\underline{th}}$ ridge estimate in Table 3 at k = 0.03. The interpretation is complicated somewhat because the standardized coefficients do not seem to average out across all experiments in the same straightforward manner as for the non-standardized coefficients, such as those presented in Table 2. Nevertheless, the average or expected values for the standardized coefficients should be fairly close to the expected values of the non-standardized coefficients times $(\sqrt{\sum x_j^2} \div 322.375)$.

respectively. Thus, the effect of $P^2$ and $N^2$ is to force a negative sign on the ridge estimate of $\beta_5$ in the full ridge regression model.

The auxiliary ridge estimates at k = 0.03 for the other explanatory variables were the following, where -1.0 below a given explanatory variable indicates that this explanatory variable is the dependent variable regressed on the other four remaining explanatory variables:

| P | N | $P^2$ | $N^2$ | PN |
|---|---|---|---|---|
| -1.0 | 0.5502 | 0.6171 | -0.3020 | 0.1154 |
| 0.5201 | -1.0 | -0.3160 | 0.6152 | 0.1588 |
| 0.5347 | -0.2896 | -1.0 | -0.0648 | 0.7723 |
| -0.2558 | 0.5512 | -0.0634 | -1.0 | 0.7246 |

The above auxiliary ridge relationships among the explanatory variables indicate that P and N should be relatively less biased if one assumes that the true $\beta$ values are all about the same magnitude, but positive for P, N, and PN, and negative for $P^2$ and $N^2$. From earlier Equation (3), the main bias for the ridge estimate of P at k = 0.03 would be from the effect of the negative $\beta$ value for $P^2$. The other ridge estimates are such that the bias should be small. Similarly, for N, only the auxiliary ridge coefficient for $N^2$, 0.6152, is of the wrong sign for minimum bias. However, for $P^2$, the bias would have been less if the auxiliary ridge coefficients for P and PN were of negative sign. Similarly, the bias of $N^2$ would have been smaller if the coefficients for N and PN had been negative. Also, a positive sign for the auxiliary ridge coefficient of $P^2$ would have lessened bias for the ridge estimate of the $N^2$ coefficient in the full model.

Although the preceding examination of the interrelationships among the explanatory variables indicates that the model of (15) is not ideal for ridge estimation, the ridge estimates of Table 3 are, nevertheless, lower in MSE than the OLS estimates in (15). Sum of the variances for the OLS estimates gives MSE $\doteq$ 5.536 in Table 3. Although the ridge estimates had MSE about 60 percent of that for OLS, provided that $k \leq 0.10$, a much smaller bias, but larger variance, was obtained from partitioned ridge regression, as shown in the next section.

## Estimation of the Production Function Coefficients by Partitioned Ridge Regression

According to prior information based upon other fertilizer experiments, a positive coefficient would be expected for the PN interaction variable of (15). Based simply upon the principle of diminishing returns, positive coefficients would be expected for the linear N and P variables and negative signs for the coefficients of the two squared terms, $N^2$ and $P^2$.

Given these expected signs and the interrelationships among the explanatory variables, a partition of the three variables with expected positive coefficients, P, N, and PN, into one block of variables, $\bar{X}_1$, would seem promising. Similarly, the two squared terms, $P^2$ and $N^2$, should be partitioned into the second block, $\bar{X}_2$. Following this procedure, the partitioned ridge estimates were obtained and are shown in Table 4. However, before examining the results of Table 4, it should again be emphasized that the model of (14) estimated in Table 4 was an unfavorable situation for partitioned ridge regression because of the relatively high correlation between variables of different blocks, as shown below:

| | P | N | PN | $P^2$ | $N^2$ |
|---|---|---|---|---|---|
| BLOCK 1 | 1.0 | 0.911 | 0.940 | 0.957 | 0.859 |
| | | 1.0 | 0.941 | 0.853 | 0.962 |
| | | | 1.0 | 0.963 | 0.963 |
| BLOCK 2 | | | | 1.0 | 0.858 |
| | | | | | 1.0 |

As hypothesized in an earlier section, smaller MSE was obtained in Table 4 at larger k values for partitioned ridge regression, as compared to regular ridge regression in Table 3. Smallest MSE = 1.8424 was obtained at k = 0.002 in Table 3, but then MSE rose rapidly for larger values of k. By contrast, in Table 4 MSE continued to decline as k was increased, reaching the lowest point, MSE = 2.2940, for k = 1.00, the largest value of k presented in Table 4.[7/]

[7/] Slightly smaller MSE = 1.1063 for the estimated coefficients of $P^2$ and $N^2$ were obtained at k = 1.20.

Table 4. Standardized Partitioned Ridge Regression Estimates Corresponding to Expected Values of Non-standardized Estimates of Iowa Corn Production Function Coefficients, Along with Estimated Variance and MSE for Various k Values

| Type of estimate | Explanatory variable Block #1 | | | Explanatory variable Block #2 | | Sum[b/] |
|---|---|---|---|---|---|---|
| | P | N | PN | $P^2$ | $N^2$ | |
| Coefficient @ k = 0.10... | 1.0984 | 1.0386 | 0.8340 | -1.1567 | -1.0822 | -- |
| Variance, k = 0.10..... | 0.2108 | 0.1718 | 0.9803 | 0.8888 | 0.8043 | 3.0560 |
| Bias squared[a/] ..... | 0.0534 | 0.0279 | 0.0847 | 0.0075 | 0.0012 | 0.1747 |
| Estimated MSE..... | 0.2643 | 0.1997 | 1.0650 | 0.8963 | 0.8054 | 3.2307 |
| Coefficient @ k = 0.20... | 1.0230 | 0.9838 | 0.8665 | -1.0914 | -1.0387 | -- |
| Variance, k = 0.20..... | 0.2378 | 0.2081 | 0.6625 | 0.7825 | 0.7256 | 2.6164 |
| Bias squared[a/] ..... | 0.0940 | 0.0492 | 0.1047 | 0.0230 | 0.0060 | 0.2769 |
| Estimated MSE..... | 0.3317 | 0.2573 | 0.7672 | 0.8055 | 0.7316 | 2.8933 |
| Coefficient @ k = 0.60... | 0.8710 | 0.8546 | 0.8151 | -0.9039 | -0.8796 | -- |
| Variance, k = 0.60..... | 0.2231 | 0.2105 | 0.3636 | 0.5322 | 0.5102 | 1.8396 |
| Bias squared[a/] ..... | 0.2103 | 0.1232 | 0.0741 | 0.1151 | 0.0560 | 0.5787 |
| Estimated MSE..... | 0.4334 | 0.3338 | 0.4376 | 0.6472 | 0.5662 | 2.4183 |
| Coefficient @ k = 1.00... | 0.7733 | 0.7630 | 0.7411 | -0.7748 | -0.7590 | -- |
| Variance, k = 1.00..... | 0.1875 | 0.1801 | 0.2659 | 0.3906 | 0.3783 | 1.4024 |
| Bias squared[a/] ..... | 0.3094 | 0.1960 | 0.0393 | 0.2193 | 0.1276 | 0.8916 |
| Estimated MSE..... | 0.4969 | 0.3761 | 0.3051 | 0.6099 | 0.5060 | 2.2940 |

[a/] Estimates of bias squared are based upon assumption that coefficients in (14), estimated from all 114 observations, are the true parameters.

[b/] Sums may not check exactly, due to rounding.

Smaller MSE was obtained for regular ridge regression in Table 3 for small values of k than for partitioned ridge regression in Table 4. However, as noted earlier in discussing the results of Table 3, when the true coefficients are unknown there is no precise method for estimating the optimal value of k to select. At best, k = 0.03 might be selected most of the time, using one of the more conservative rules for selection of k.[8/]

If values of k of 0.1 or larger are used, then smaller MSE values were obtained with partitioned ridge regression than for regular ridge regression. For k = 0.1, MSE $\doteq$ 3.31 in Table 3 versus MSE $\doteq$ 3.23 in Table 4 for k = 0.1. If k = 0.20 were always used, then MSE $\doteq$ 4.13 would be obtained by regular ridge regression (not shown in Table 3). However, the partitioned ridge estimate has MSE $\doteq$ 2.89 at k = 0.20 in Table 4, only about 70 percent as high as for regular ridge regression.

Although not lower in MSE for small values of k, two significant advantages of partitioned ridge regression for the model of (15) and Tables 3 and 4 should be noted:

1. The bias squared component of MSE was much smaller for partitioned ridge regression in Table 4 than for regular ridge regression in Table 3. This is an important advantage since a fairly precise and unbiased estimate of the variance component of MSE can be obtained directly from the sample data.[9/] For example, for

---

[8/] This statement is somewhat speculative without running a thorough Monte Carlo study of the model. However, in unpublished Monte Carlo studies by the author with other models, none of the presently suggested methods for selecting k were very accurate in selecting optimal k values.

[9/] Admittedly it is possible to estimate MSE of the ridge estimator directly from the sample data, but such an estimate may be highly variable. It is simple but tedious to show, using expected values and properties of the matrix trace, that the expected MSE of the standardized ridge estimator is:

$$(18)\quad E(\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta) = E(\hat{\beta}^* - \hat{\beta})'(\hat{\beta}^* - \hat{\beta}) + \sigma^2 \operatorname{tr}[2(X'X + kI)^{-1} - (X'X)^{-1}].$$

Equation (18) follows from the fact that

(continued on following page)

k = 0.1, the bias squared component is over 99 percent of the MSE of the regular ridge estimate in Table 3. On the other hand, bias squared is only about 5 percent of the MSE of the partitioned ridge estimator in Table 4. Even for k = 1.0 in Table 4, bias squared is less than 39 percent of the MSE. Thus, the researcher would be far less apt to be misled by the partitioned ridge estimates, compared to the regular ridge estimates.

2. Another important advantage of the partitioned ridge estimates of Table 4 over the regular ridge estimates of Table 3 is that the partitioned estimates are not nearly as sensitive to the value of k selected. For example, in increasing k from 0.20 to 1.00 in Table 4, MSE decreases from 2.8933 to 2.2940, a change of $0.5993 \div 2.8933 \doteq 21$ percent. By contrast, if k were increased from 0.03 to 0.40 for regular ridge regression, MSE would be increased from 2.3651 (shown in Table 3) to 4.8968 (not given in Table 3), or a more than doubling of MSE. The point is that the researcher should have a much better chance of selecting a satisfactory value of k for the model of Tables 3 and 4 with partitioned ridge regression than for regular ridge regression, especially considering that variance would be a much larger component of MSE, as discussed under Item 1 above.

---

(Footnote 9 continued):

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

and that

$$E(\hat{\beta}^* - \hat{\beta})'(\hat{\beta} - \beta) = tr[E(\hat{\beta} - \beta)(\hat{\beta}^* - \hat{\beta})']$$
$$= tr\{E[(X'X)^{-1} X'u]\}\{[X'X + kI)^{-1} - (X'X)^{-1}] X'(X\beta + u)\}'$$
$$= \sigma^2 tr[(X'X + kI)^{-1} - (X'X)^{-1}] \quad \text{since } E(u) = 0 \text{ and } E(uu') = \sigma^2 I_n.$$

## Supplementation of Partitioned Ridge Regression with the Mean Vector

Given that a ridge regression model can be partitioned into two or more "compatible blocks" of variables,[10/] depending upon the expected signs of the model parameters and the direction of interrelationships among the explanatory variables, then the researcher can go a step further and supplement the partitioned ridge estimates with the mean vector of each block, similar to the procedure used for the models of Table 1, combined with the partitioned ridge regression procedure. To illustrate these remarks, suppose that we have two compatible blocks of explanatory variables. Then, again use OLS to fit the full model, $\hat{Y} = \bar{X}_1\hat{\beta}_1 + \bar{X}_2\hat{\beta}_2$. Also, define $Z_1 = Y - \bar{X}_2\hat{\beta}_2$ and $Z_2 = Y - \bar{X}_1\hat{\beta}_1$, just as before for partitioned ridge regression. But instead of fitting these two partitioned models by regular ridge regression, the ridge estimate of $Z_1 = \bar{X}_1\beta_1$ can be supplemented by the mean of the OLS estimates of the q coefficients of $\bar{X}_1$. That is, let

$$(19) \qquad \tilde{b}_1 = [(\bar{X}_1'\bar{X}_1 + k\lambda_1)^{-1}(I_q + k\lambda_1 R)]\bar{X}_1'z_1 = A\bar{X}_1'z_1$$

$$= A\bar{X}_1'(Y - \bar{X}_2\hat{\beta}_2) = A\bar{X}_1'[\bar{X}_1\beta_1 - \bar{X}_2(\hat{\beta}_2 - \beta_2)].$$

In (19), all the symbols are the same as defined earlier in (5b) and (8). Following the same procedure as used earlier for Equations (9) through (12), it is not difficult to derive $\mathrm{Var}(\tilde{b}_1)$ in (20), assuming fixed X values, $E(u) = 0$, and $E(uu') = \sigma^2 I_n$:

$$(20) \qquad \mathrm{Var}(\tilde{b}_1) = \sigma^2 A\bar{X}_1'\bar{X}_1A' + A\bar{X}_1'\bar{X}_2[\mathrm{Var}(\hat{\beta}_2)]\bar{X}_2'\bar{X}_1A'.$$

It is apparent that $\mathrm{Var}(\tilde{b}_1)$ in (20) is of the same form as the variance of the partitioned ridge estimate given earlier in (12), although the matrix $A \neq (X'X + k\lambda_1)^{-1}$, except when $k = 0$. (Of course, when $k = 0$, $b_1^* = \tilde{b}_1 = \hat{\beta}_1$.)

---

[10/] Perhaps it should be noted explicitly that a "compatible block" should, ideally, be constructed so that each explanatory variable would be a positive function of those other variables within the block which have the same expected coefficient sign and be a negative function of those variables within the block which have an opposite expected coefficient sign.

Also, just as in comparing the regular ridge estimates with the OLS-mean-vector supplemented ridge estimates in Table 1, the variance of $\tilde{b}_1$ usually exceeds that of $b_1^*$.

For the earlier model of (13) and Table 2, the expected values, variances, and MSE of $\tilde{b}_1$ and $\tilde{b}_2$ for various k values were computed:

| Value of k | $E(\tilde{b}_1)$ | $E(\tilde{b}_2)$ | $V(\tilde{b}_1) + V(\tilde{b}_2)$ | $MSE(\tilde{b}_1) + MSE(\tilde{b}_2)$ |
|---|---|---|---|---|
| 0.0 (OLS) | 3.000 | 2.000 | 21.667 | 21.667 |
| 0.50 | 2.519 | 2.481 | 2.695 | 3.157 |
| 1.00 | 2.510 | 2.490 | 2.674 | 3.155 |
| 10.00 | 2.501 | 2.499 | 2.667 | 3.165 |

Although the above mean-vector-supplemented partitioned ridge estimates are slightly higher in MSE at k = 0.50 than for the unsupplemented partitioned ridge estimate (which had $MSE(b_1^*) = 2.698$ at k = 0.50), it should be observed that the bias squared component of $MSE(\tilde{b}_1)$ is only about one-half as large at k = 0.50. Along these same lines, if the variance error term, $u_i$, of (13) had been one-fourth as large, that is if $E(u_i) = 0.5(-2) + 0.5(2)$ and $E(u_i^2) = 4$, then a lower MSE would be obtained from the mean-supplemented partitioned ridge estimator at k = 0.50. In that case we would have, for k = 0.50,

$$MSE(b_1^*) = V(b_1^*) + Bias^2(b_1^*) \doteq 0.432 + 0.971 \doteq 1.403, \quad \text{and}$$

$$MSE(\tilde{b}_1) = V(\tilde{b}_1) + Bias^2(\tilde{b}_1) \doteq 0.674 + 0.462 \doteq 1.136.$$

Thus, MSE for the mean-supplemented partitioned ridge estimator becomes relatively more favorable when $\sigma^2$ and the coefficient variances are smaller relative to the bias squared component of MSE. Although more research is obviously needed, the mean-supplemented partitioned ridge estimator appears especially promising for those economic (and other) models with a high $R^2$ value for the full regression model. One other interesting feature of the mean-supplemented partitioned ridge estimator is that the minimum variances for the coefficient estimates should

usually be obtained by letting A = R in (20). (It is apparent from (5a) that b tends to $\hat{\beta}$ as k becomes large.)

## SUMMARY AND CONCLUSIONS

Various alternatives for dealing with the problem of multicollinearity have been explored in this paper. The first method studied was the supplementation of ridge regression with prior information. Although intuitively appealing, this type of estimator gave results somewhat less reliable than expected when the prior information was related to the sample data.

Given the difficulties encountered with supplementation of ridge regression with a sample-related prior vector, a "partitioned" ridge procedure was developed and applied to a simple model by means of a simple Monte Carlo study. Mean square error from partitioned ridge regression was about one-half that from OLS and regular ridge regression.

Regular and partitioned ridge regression estimates were then obtained for a production (yield-fertilizer response) function fitted to a nonorthogonal subset of data from an earlier study [Heady, et al., 1955]. It should be noted that the production function example was poor from the viewpoint of both regular and partitioned ridge regression. The interrelationships among the explanatory variables and the true β coefficients caused fairly high bias for the regular ridge estimator, as discussed in more detail earlier. On the other hand, the partitioned ridge estimator had much lower bias, but fairly high variance, the high variance resulting from the high intercorrelation between variables of different blocks. Much lower variance could have been obtained if the highest correlations had been between the variables within blocks rather than between variables located in different blocks.

Despite these limitations for both types of ridge regressions, both the regular and partitioned ridge estimates had about one-half the MSE of OLS over a considerable range of k values. However, given the wider range of k values over which a low MSE could be obtained, the partitioned ridge estimator was thought to

be more reliable than the regular ridge estimator, especially considering that the easily measured variance was the major component of MSE for the partitioned ridge estimator, whereas the usually unknown bias squared was the major part of MSE for the regular ridge estimator.

Given the results from fitting the earlier artificial models and the empirical production function of Tables 3 and 4, the following procedure is recommended:

1. Start by fitting the fully specified model by OLS. Examine the t values and signs of the OLS-estimated coefficients. If some signs are illogical and unexpected and t values are relatively low for some of the coefficients (but with a high level of significance for the overall regression), then examine the VIF (main diagonal elements of the inverted correlation matrix) for further evidence of multicollinearity. If the VIF are quite high for some of the variables with low t values and/or "wrong" signs, then regular ridge regression may be appropriate.[11/]

2. In fitting regular ridge regression, the procedure should not be blindly followed. Rather, the interrelationships among the explanatory variables, Equation (3), should be examined in conjunction with the postulated signs of the regression coefficients. If the interrelationships and expected signs of the regression coefficients appear to be fairly compatible in terms of earlier Equation (3), then some faith in the regular ridge estimates is probably justified. However, the problem of which k value to use still remains. Some insight may be gained by using Wallace's [1972] u statistic. Similarly, criteria used by Mallows [1973] or Farebrother [1975] may be helpful in selecting a k value.

---

11/ In this discussion it is assumed that the model was properly specified at the start in the sense that no important, relevant explanatory variables have been excluded. If so, then a low t value and a relatively small VIF for a given explanatory variable constitute evidence that the variable should be deleted. However, if the low t value results from a large VIF, then there is much less justification for deletion.

3. If the interrelationships among the explanatory variables and expected signs of the coefficients appear to combine to give large biases in terms of (3), then some other alternative, such as partitioned ridge regression, should be explored. It may be possible, as illustrated by the three models fitted in this paper, to partition the regression model into two or more "compatible" blocks, thus significantly reducing unwanted bias. Admittedly, the reduction in bias is somewhat offset by an increase in variance, but the increased variance has been small compared to that from OLS in the models fitted thus far.

4. An even further reduction in MSE can sometimes be obtained by supplementing the partitioned ridge estimator with the OLS mean vector. This estimator has the advantage of tending to the mean vector as k becomes large, rather than tending toward zero as for the unsupplemented ridge estimators. Although the mean-vector supplemented partitioned ridge estimator gave smaller bias for the example fitted, some of this advantage was offset by an increase in variance.

Although more research is obviously needed, the partitioned ridge approach presented in this paper should substantially increase the applicability and accuracy of biased linear estimation for many regression problems. One reason for this alleged increased applicability and accuracy is that the partitioned approach makes use of the interrelationships existing among the explanatory variables and uses prior knowledge about the expected signs of the regression coefficients. It is interesting that there is very little cost for utilizing information about the interrelationships among the explanatory variables; these relationships are already inherent within the sample data. Similarly, the researcher should usually know the direction of the effect of a particular explanatory variable; otherwise, he is hardly competent to specify the regression model in the first place.

REFERENCES

[1] Brown, W. G., Effect of Omitting Relevant Variables Versus Use of Ridge Regression in Economic Research. Oregon Agr. Exp. Sta. Special Report 394, Corvallis, 140 pp. October 1973.

[2] Brown, W. G. and Bruce R. Beattie, "Improving Estimates of Economic Parameters by Use of Ridge Regression With Production Function Applications," Am. J. Agr. Econ., 57:21-32, February 1975.

[3] Burt, Oscar R., "Multicollinearity and Elliptically Constrained Parameter Vectors in Regression," paper submitted for publication in J. Am. Stat. Ass'n.

[4] Farebrother, R. W., "The Minimum Mean Square Error Linear Estimator and Ridge Regression," Technometrics, 17:127-128, February 1975.

[5] Goodnight, James, and T. D. Wallace, "Operational Techniques and Tables for Making Weak MSE Tests for Restrictions in Regression," Econometrica, 40:699-709, July 1972.

[6] Heady, Earl O., John T. Pesek, and W. G. Brown, Crop Response Surfaces and Economic Optima in Fertilizer Use, Iowa Agr. Exp. Sta. Res. Bul. 424, Ames, March 1955.

[7] Hoerl, A. E. and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, 12/1:55-67, February 1970.

[8] _____, "Ridge Regression: Applications to Nonorthogonal Problems." Technometrics, 12/1:69-82, February 1970.

[9] Mallows, C. L., "Some Comments on $C_p$," Technometrics, 15/4:661-675, November 1973.

[10] McDonald, G. C. and R. C. Schwing, "Instabilities of Regression Estimates Relating Air Pollution to Mortality," Technometrics, 15:463-481, August 1973.

[11] Swindel, Benee F., "Good Ridge Estimators Based on Prior Information," paper presented at annual meeting of the American Statistical Association and the Biometric Society, St. Louis, August 26-29, 1974.

[12] Theil, H. and A. S. Goldberger, "On Pure and Mixed Statistical Estimation in Economics," Int. Econ. Rev. 2:65-78. 1961.

[13] Wallace, T. D., "Weaker Criteria and Tests for Linear Restrictions in Regression," Econometrica, 40:689-698, July 1972.

# APPENDIX

Table A-1. Subset of Data Analyzed for Production Function Estimation[a/]

| Pounds $P_2O_5$ per acre | Pounds nitrogen per acre | Bushels corn per acre | Pounds $P_2O_5$ per acre | Pounds nitrogen per acre | Bushels corn per acre |
|---|---|---|---|---|---|
| 0 | 0 | 24.5 | 200 | 160 | 109.3 |
| 0 | 0 | 6.2 | 160 | 200 | 105.7 |
| 40 | 0 | 26.7 | 160 | 200 | 115.5 |
| 40 | 0 | 29.6 | 200 | 200 | 140.3 |
| 0 | 40 | 23.9 | 200 | 200 | 142.2 |
| 0 | 40 | 11.8 | 160 | 240 | 130.5 |
| 40 | 40 | 60.2 | 160 | 240 | 124.3 |
| 40 | 40 | 82.5 | 240 | 240 | 121.1 |
| 80 | 80 | 99.5 | 240 | 240 | 114.2 |
| 80 | 80 | 115.4 | 320 | 240 | 127.3 |
| 160 | 80 | 102.2 | 320 | 240 | 139.5 |
| 160 | 80 | 108.5 | 280 | 280 | 130.0 |
| 120 | 120 | 119.4 | 280 | 280 | 141.9 |
| 120 | 120 | 97.3 | 320 | 280 | 131.8 |
| 160 | 120 | 133.3 | 320 | 280 | 111.9 |
| 160 | 120 | 124.4 | 240 | 320 | 130.9 |
| 120 | 160 | 113.6 | 240 | 320 | 144.9 |
| 120 | 160 | 102.1 | 280 | 320 | 124.8 |
| 160 | 160 | 129.7 | 280 | 320 | 114.1 |
| 160 | 160 | 116.3 | 320 | 320 | 127.9 |
| 200 | 160 | 128.7 | 320 | 320 | 118.8 |

[a/] Taken from Heady et al., 1955, p. 330.