R-716-PR

April 1971

AN EVALUATION OF RIDGE ESTIMATORS

Joseph P. Newhouse and Samuel D. Oman

A Report prepared for

UNITED STATES AIR FORCE PROJECT RAND

Rand
SANTA MONICA, CA. 90406

This research is supported by the United States Air Force under Project Rand—Contract No. F44620-67-C-0045—Monitored by the Directorate of Operational Requirements and Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this study should not be interpreted as representing the official opinion or policy of Rand or of the United States Air Force.

R-716-PR
April 1971

AN EVALUATION OF RIDGE ESTIMATORS

Joseph P. Newhouse and Samuel D. Oman

A Report prepared for

UNITED STATES AIR FORCE PROJECT RAND

Rand
SANTA MONICA, CA. 90406

Published by The Rand Corporation

PREFACE

This Memorandum is part of Rand's effort to improve the techniques available to analyze non-experimentally generated data. A particularly useful statistical tool for this purpose is multiple regression analysis, which can be used to study the effect on a particular variable of changing one explanatory variable while holding the others constant. In a particular sample, however, explanatory variables may be interrelated. Then ordinary regression techniques do not give precise estimates of the effect of changing any particular explanatory variables. This Memorandum evaluates an alternative to the ordinary technique that has been claimed to be useful in such a situation.

Interrelatedness of observed explanatory variables (or collinearity, as it is known technically) occurs all too frequently in non-experimental data. For example, in estimating the budgetary cost of a volunteer army of various sizes, one needs to know the effect of relative military-civilian wages on the enlistment rate, holding the unemployment rate constant. Data from various regions at one point in time have been used to estimate this effect. However, if high unemployment occurs in the same areas as high ratios of military to civilian pay, it is hard to estimate the effect of adjusting only the military to civilian pay ratio.

Another example of the problem is estimating the supply of volunteers to the Air Force as a function of first term wages and career wages. Historically, first-term and career wages have tended to move together. As a result, it is difficult to estimate with accuracy what the effect would be of changing, say, first-term wages while keeping career wages constant. Rand studies that discuss these two particular problems are S. L. Canby and B. P. Klotz, The Budget Cost of a Volunteer Military, RM-6184-PR, August 1970; and A. A. Cook, Jr., The Supply of Air Force Volunteers, RM-6361-PR, September 1970.

The ridge analysis technique evaluated in this study purports to be able to determine more precisely the effects of changes in various explanatory variables. Since this technique has received attention in the literature, and since the problem it purports to solve occurs frequently, it was thought that an analysis of its properties would be useful. We found that the estimator as proposed contains serious flaws and we caution against the use of ridge analysis, as now defined, to estimate regression coefficients.

SUMMARY

This Memorandum explores the properties of a class of estimators, variously known as ridge analysis or ridge regression estimators, proposed as an alternative to ordinary least squares regression when sample data are collinear. A ridge estimator has been defined as the estimate that minimizes the residual sum of squared deviations, subject to the constraint that the length of the estimated coefficient vector be less than or equal to r, where r is less than the length of the ordinary least squares estimate. It has been asserted that there is a rule for choosing r such that the corresponding ridge estimator does better than ordinary least squares with respect to the unweighted sum of coefficient mean squared errors; however, this assertion has not been proved. Rather, various _ad hoc_ procedures for selecting r have been proposed. Using Monte Carlo techniques, we have evaluated procedures suggested in the literature; we have also tried a number of variants.

All the ridge estimators proposed did worse than ordinary least squares for at least some choices of the true regression coefficients. Moreover, in our opinion, the ridge estimators failed by a sufficient margin and in a sufficient number of cases to preclude their use.

However, the performance of the ridge estimators was correlated with the sample multiple correlation coefficient in the models considered. It thus appears that it might be possible to define a ridge estimator that would be better than ordinary least squares. Until the properties of such an estimator are rigorously derived, however, we caution the reader against using ridge analysis to estimate regression coefficients.

CONTENTS

I. INTRODUCTION

In applications of multiple linear regression it usually occurs that the explanatory variables under consideration are not all orthogonal; that is, their sample correlation matrix departs from the identity matrix. If the departure is sufficiently large, the ordinary least squares (OLS) estimate of the regression coefficients tends to become "unstable"; several of the estimated coefficients will often be inordinately large in absolute value and have large standard errors.

Ridge analysis (or ridge regression) has been proposed by A. E. Hoerl as an alternative to least squares estimation when dealing with such multicollinearity. Specifically, in several articles Hoerl (1959) and (1962) and Hoerl and Kennard (1970a) and (1970b) have developed ridge analysis techniques that are asserted both to portray the departure of the data from orthogonality and to result in coefficient estimates "better" than the least squares estimates (with respect to mean squared error). However, this second property, which is our concern here, has not been mathematically proved in any of these articles.

Since ridge analysis is an ad hoc procedure, a logical preliminary estimate of its effectiveness would be its empirical performance. This is attempted in Hoerl and Kennard (1970b); however, since the true values of the parameters being estimated are not given, we do not believe the statement made on p. 72 of that article is justified, that ridge coefficients "will undoubtably be closer to [the true parameters] and are more stable for prediction than the least squares coefficients."

We have conducted a series of Monte Carlo experiments to compare the performance of ridge analysis with least squares. Our essential results are twofold: First, all the ridge estimators proposed by Hoerl and Kennard did worse than OLS for at least some (non-pathological) choices of the true regression coefficients. Moreover, in our opinion, the ridge estimators failed by a sufficient margin and in a sufficient number of cases to preclude their use. Second, the performance of the ridge estimators was correlated with the sample multiple correlation coefficient (R^2) of the samples considered. It thus appears that it might be possible to define a ridge estimator that would be better than OLS. Until the properties of such an estimator are rigorously derived, however, we would caution the reader against using ridge analysis to estimate regression coefficients.

II. THEORETICAL CONSIDERATIONS

Consider the standard multiple linear regression model

$$y = X\beta + \varepsilon , \tag{2.1}$$

where $X = (x_{ij})$ is a fixed $N \times n$ matrix (x_{ij} is the ith observation on the jth independent variable),

$y = (y_i)$ is an $N \times 1$ vector of observations on the dependent variable,

and $\varepsilon = (\varepsilon_i)$ is an $N \times 1$ random vector which is distributed $\text{Normal}(0,\sigma^2 I)$.

We wish to estimate the $n \times 1$ vector β of regression coefficients.

Let us assume the variables have been standardized by subtracting their sample means and dividing by their sample standard deviations.

If β^* is an estimator of β, we take as a measure of its goodness its unweighted sum of mean squared errors

$$R(\beta,\beta^*) = E[L(\beta,\beta^*)] = E[(\beta^* - \beta)'(\beta^* - \beta)] = \sum_{i=1}^{n} E(\beta_i^* - \beta_i)^2. \tag{2.2}$$

Hoerl and Kennard have defined ridge estimators as follows: Let T be an $(n \times n)$ orthogonal matrix ($T'T = I$) such that $T'ST = D$, where $S = X'X$ is the correlation matrix of the independent variables and D is the diagonal matrix whose ith element is the ith eigenvalue λ_i of S. (We shall suppose $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$.) By transforming (2.1) to the orthogonal regression model

$$y = \tilde{X}\tilde{\beta} + \varepsilon , \tag{2.3}$$

where $\tilde{X} = XT$ and $\tilde{\beta} = T^{-1}\beta$, the correlation matrix $S = X'X$ is transformed to the covariance matrix $\tilde{S} = \tilde{X}'\tilde{X} = (T'X')(XT) = T'(X'X)T = D$. Now let K be an $(n \times n)$ diagonal matrix with non-negative elements; the ridge estimate $\tilde{\beta}_K$ for the (transformed) model (2.3) has been defined by

$$\tilde{\beta}_K = (D + K)^{-1}\tilde{X}'y , \tag{2.4}$$

and the ridge estimate β_K for the (original) model (2.1) by

$$\beta_K = T\tilde{\beta}_K \ . \tag{2.5}$$

(Actually, the above definitions define a class of estimators, indexed by the matrix function K. A <u>bona fide</u> estimator is defined only when K is defined; that is, only when the rule for determining "which K to use" is specified. Indeed, the crux of this study is the problem of specifying K. (2.4) and (2.5) should thus be taken as definitions "modulo K".)

If $K = kI$ is a scalar matrix, we shall write β_K as β_k. Note that if $K \equiv 0$, then $\tilde{\beta}_0$ is the OLS estimate for the model (2.3), and β_0 is the OLS estimate for our original problem. Also note that if $K = kI$, then

$$\begin{aligned} \beta_k &= T\tilde{\beta}_k \\ &= T(D + K)^{-1}\tilde{X}'y \\ &= T[T'(S + K)T]^{-1}(T'X'y) \\ &= (S + K)^{-1}X'y \ . \end{aligned} \tag{2.6}$$

Hoerl and Kennard (1970a) have shown several geometric properties to hold for ridge estimators; these properties are discussed in Appendix B. The following two facts were proved in Hoerl and Kennard (1970a).

If $K = (k_i)$ is constant, then

Fact 1: $$R(\beta,\beta_K) = \sigma^2 \sum_{i=1}^{n} \frac{\lambda_i}{(\lambda_i + k_i)^2} + \sum_{i=1}^{n} \frac{k_i^2 \tilde{\beta}_i^2}{(\lambda_i + k_i)^2} \ . \tag{2.7}$$

Fact 2: If $0 < k_i \le \frac{\sigma^2}{\tilde{\beta}_i^2}$ for all i, with $k_i < \frac{\sigma^2}{\tilde{\beta}_i^2}$ for at least one i, then

$$R(\beta,\beta_K) < R(\beta,\beta_0) \ . \tag{2.8}$$

In other words, for a given model (2.1) there exists a constant matrix K such that β_K is a better estimate than β_0 of β. Unfortunately, the matrix K depends on parameters σ^2 and $\tilde{\beta}$ which are unknown to the investigator. Hoerl and Kennard (1970a) recognize and discuss this problem to an extent. They then assert (p. 64), however, that "it is not difficult to select a β_K that is better than β_0," and propose several methods for choosing the associated K.

In all these methods, the choice of K depends on the particular sample (observation vector y) being considered. But then the assumption that K is a fixed matrix, independent of the random vector ε, is violated. This in turn invalidates their proof that β_K is better than β_0, since their derivation of (2.7), from which (2.8) follows immediately, depends heavily on the assumed independence of K. Although the authors do not state that (2.7) is still valid when K is variable, nonetheless they do not, we believe, sufficiently emphasize the fact that it is not.

This of course does not mean that the ridge estimators proposed in Hoerl and Kennard (1970a) will <u>necessarily not</u> be better than OLS; it simply means that they will <u>not necessarily</u> be better than OLS. We therefore conducted the Monte Carlo experiments described in Section III; our results indicate that the preceding remarks have more than simply theoretical implications.

III. THE MONTE CARLO EXPERIMENTS

The Monte Carlo experiments all dealt with the dimension $n = 2$. N (=100) observations of the independent variables (X_1 and X_2) were generated, such that their sample correlation was r; this was done for $r = .9$ and $r = .99$. The observations were computed in the following manner: Three independent sequences (U, V_1, V_2) of independent Normal (0,1) pseudo-random numbers, each of length N, were generated. X_i was then set to $U + \alpha V_i$, where α was chosen to result in the desired sample correlation between X_1 and X_2. Since

$$r = \frac{\sum_i x_{i1}x_{i2}}{(\sum_i x_{i1}^2)^{\frac{1}{2}}(\sum_i x_{i2}^2)^{\frac{1}{2}}} \quad , \alpha \approxeq [(1-r)/r]^{\frac{1}{2}} \quad .$$

The sample means were subtracted from the X_i, but they were not divided by their sample variances; therefore in the following discussion $S = X'X$ will denote the matrix of cross-products.

For each S so constructed, a set of seven true coefficient vectors $\beta^1, \beta^2, \ldots, \beta^7$ were considered. The β^j were all of unit length, but chosen to make different angles with the normalized eigenvector v_1, corresponding to the eigenvalue λ_1, of the matrix S. Denoting the angle between β^j and v_1 by θ_j, we have $\theta_1 = 0°$ (that is, $\beta^1 = v_1$), $\theta_2 = 15°$, $\theta_3 = 30°$, ..., $\theta_7 = 90°$ (that is, $\beta^7 = v_2$, the eigenvector corresponding to the eigenvalue λ_2). We considered these values of β^j for the following reason: As remarked earlier, β_K is the OLS estimate when $K \equiv 0$, so in this case (2.7) reduces to $R(\beta,\beta_0) = \sigma^2 \sum_{i=1}^{n} \frac{1}{\lambda_i}$. From this we see that the OLS mean squared error depends only on the independent variables (through the λ_i) and the error variance σ^2; it does not depend on the true coefficient vector β. When K is a constant nonzero matrix, however, the second sum in (2.7) shows that for fixed λ_i and σ^2, $R(\beta,\beta_K)$ depends additionally on the true coefficient vector

β (or equivalently, on $\tilde{\beta}$). It is clear, for example, that if $\gamma = 2\beta$ is another coefficient vector, then the second sum in (2.7) will be four times as large for γ as for β; thus $R(\beta, \beta_K)$ depends on the length of β. Moreover, if the ratios $\frac{k_i}{\lambda_i}$ are not equal for all i, then (2.7) shows that $R(\beta, \beta_K)$ depends also on the relative sizes of the components β_i of the vector β. For example, if $\lambda_1 > \lambda_2 > \dots > \lambda_n$, then by letting $K \equiv kI$ and regarding $R(\beta, \beta_K)$ as a function of β, it is easy to show that R is minimized, subject to $||\beta|| = 1$, by $\beta = v_1$; R is maximized when $\beta = v_n$.

Although these remarks do not prove anything (since (2.7) is not necessarily true when K is not constant), they do indicate that a preliminary empirical evaluation of ridge estimators should take the relative sizes of the components of β into account.

For each S and β^j, M(=100 or 1000) trial cases (M observed vectors y) were generated according to (2.1), where σ^2 was held constant at 1/9. This value was chosen so that the multiple correlation coefficient would be above .9 for at least some β^j.

Thus for a given matrix X, each estimator was used in M trials to estimate seven true coefficient vectors β^j. In the tables found in Appendix A, for each β^j the mean squared error (MSE) for each component of the vector β^j, averaged over the M trials, is presented for each estimator.[1]

---

[1]The values given in the tables are the MSEs for the non-standardized coefficients, that is, for the case in which we do not assume that the variables in (2.1) have been divided by their sample standard deviations. This MSE is not in general equal to the MSE for the standardized coefficients, so that one estimator having a lower non-standardized MSE than another estimator does not imply that its standardized MSE will also be lower. However, if one estimator's non-standardized MSE for both coefficients is lower, then it is easy to see that its standardized MSE will also be lower than that of the other estimator.

In some experiments several ridge estimators were tested; in all experiments the OLS estimate was computed, so that the MSE for each ridge estimator could be compared with the OLS MSE.[1]

---

[1]It might be objected that, since OLS is known to be an admissible estimator in the two dimensional case (that is, there exists no β^* such that $R(\beta,\beta^*) \leq R(\beta,\beta_0)$ for all β, with strict inequality holding for at least one β), we are not justified in comparing ridge estimators with OLS when $n = 2$. However, we point out that it is still possible that $R(\beta,\beta_K)$ might be significantly lower than $R(\beta,\beta_0)$ for some values of β, while not being too much greater for other values of β. Moreover there is nothing in the nature of the ridge estimators proposed thus far that would lead one to believe that results in higher dimensions would be significantly different from those obtained in the two-dimensional case.

IV. THE RESULTS

The experiments were done for a number of ridge estimators that seemed to fall into two main categories. The first category of estimators did much better than OLS for a narrow range of angles θ (of the true coefficient vector β with the first eigenvector of the matrix S), namely, $\theta = 0°$ when the correlation r was .90, and $\theta = 0°, 15°$ for $r = .99$. For the remaining angles, however, these estimators did markedly worse than OLS (except for $\theta = 90°$, in which case some estimators "collapsed" to OLS by selecting $K = 0$). The estimators in the second category did better than OLS in a wider range of angles for $r = .99$ ($\theta = 0°, 15°, 30°, 45°$) although their margin of improvement was not as great as with the first category. Similarly, these estimators did not do as badly for the unfavorable angles as did the first category of estimators. Nonetheless, the increased range of superiority of these estimators did not seem to outweigh the disadvantages of using them in the unfavorable cases. One exception to this classification was a ridge estimator that turned out to be virtually indistinguishable from OLS because of its criterion for selecting K.

The first category contained three estimators. R1A was constructed by standardizing the variables and choosing K to be a scalar matrix kI. The value of k was determined by regarding the residual error $f(\beta_k) = (X\beta_k - y)'(X\beta_k - y)$ as a function of $\rho = ||\beta_k||$, and then selecting that value of ρ (or equivalently, of k), for which $g = \frac{d^2(f^{1/2})}{d\rho^2}$ was maximized. Computationally, this was done by increasing k from 0 in increments of .01 until a local maximum for g was reached. If no maximum was obtained in 50 iterations, then $k = .5$ was used. This estimator was originally suggested by Hoerl (1962). Although the use of the derivative criterion is not mentioned in Hoerl and Kennard (1970a), it corresponds mathematically to the suggestion given there of choosing a value of k at which the system "stabilizes," since $\max_\rho g$ is just the "knee" of the graph of $f^{1/2}$ versus ρ, that is, the point at which $f^{1/2}$ starts to increase rapidly as a function of ρ. As can be seen from Tables 2 and 3 in Appendix A, R1A did very poorly for most values of θ; for $r = .9$ and $\theta = 45°, 60°$, g consistently failed to obtain a maximum.[1] Moreover, the rather complicated analytical form of the second derivative, from which it is not even clear that g always actually has a maximum, seems to preclude a rigorous mathematical treatment of R1A.

R1B was the same as R1A except that the variables were not first standardized, and k was increased by increments of 1 (since now S was a cross-product matrix). We considered R1B because, unlike OLS, a ridge estimator is not invariant under arbitrary non-orthogonal preliminary transformations of the data. That is, transforming to the standardized model, obtaining a ridge estimate, and then transforming back to the original model will not, in general, yield the same estimate as would be obtained without these transformations. As can be seen from Tables 2 and 3, R1B behaved essentially the same as R1A; this is because the variances of the variables X_1 and X_2 were approximately equal, so that S was approximately equal to a scalar multiple of the correlation matrix. We therefore constructed another S where var X_2 was approximately twice var X_1; the results for these experiments are shown in Table 4 in Appendix A. We see in these experiments that R1A and R1B gave markedly different estimates for the same β, but their performance averaged over all the β vectors was similar.

R1C, the procedure suggested in Hoerl and Kennard (1970a), p. 65, consisted of standardizing the variables, transforming to the orthogonal model (2.3), and then obtaining successive estimates β_{K_1}, β_{K_2}, ... until their lengths converged. For the first iteration the ith element of K_1 was set to the "optimal" value $\hat{\sigma}^2/\tilde{\beta}_{0i}^{\,2}$ where $\tilde{\beta}_0 = (\tilde{\beta}_{01}, \ldots, \tilde{\beta}_{0n})$ was the orthogonalized OLS estimate. In the experiments run, the convergence criterion was that $||\tilde{\beta}_{K_n} - \tilde{\beta}_{K_{n-1}}||^2$ be less than $.01||\tilde{\beta}_0||^2$. As can be seen from Tables 2 and 3, R1C is in no sense an optimal estimator.

[1]The similarity between the results of Table 3 with 100 trials and Table 2 with 1000 gives us confidence that, in the model considered, 100 trials produces a reasonably close estimate of the relationship of ridge analysis and OLS MSEs.

We observed that in the cases when these ridge estimators failed, they tended to require a large number of iterations, resulting in large values on the diagonal of K and correspondingly large residual sums of squares f. We therefore constructed the following R2 estimator: We first computed the R1B estimate β_k, and then took β_k as our estimate of β if $u = [f(\beta_k) - f(\beta_0)]/[f(\beta_0)/(N-n-1)]$ was less than a certain critical value u_α; if u was greater than u_α we used the OLS estimate β_0. Toro-Vizcarrondo and Wallace (1968) have shown that the statistic u, which has a non-central F-distribution, may be used to test the null hypothesis that $R(\beta,\beta^*) < R(\beta,\beta_0)$, where β^* is the OLS estimate, subject to the linear restriction that $\beta'h = c$, h being a known vector and c a known constant. For a number of reasons our use of this statistic is not rigorously justified: In particular, although a given ridge estimate β_k may be considered as the OLS estimate subject to a linear restriction (that the estimate lie in the hyperplane of vectors tangent to the ellipse $\{\beta:(\beta-\beta_0)'S(\beta-\beta_0) = \gamma(k)\}$ at the point β_k), this restriction depends on the sample at hand (through β_0 and k), violating the assumptions that h and c be prior known constants. Nevertheless, the statistic u provides a quantitative, somewhat less than arbitrary, method of measuring the increase in f over $f(\beta_0)$ due to the use of β_k; and this relative increase has been specified in Hoerl and Kennard (1970a) as one criterion for selecting k. The values for u_α were obtained by extrapolation from Wallace and Toro-Vizcarrondo (1969). From Table 5 it may be seen that for an appropriate value of α (namely, $\alpha = .50$), R2 does not have too great an increase in MSE over the unfavorable range of θ; yet its decrease in MSE over the favorable range is also tempered, so that R2 is approximately equivalent to OLS.

The last three estimators tested fell into the second category mentioned earlier. The R3A estimator was the same as R1C except that equal diagonal elements $k_i = k = \hat{\sigma}^2/||\tilde{\beta}_0||^2$ were specified for K. This value of k was chosen both because (as proved in Hoerl and Kennard (1970a)) $R(\beta,\beta_k) < R(\beta,\beta_0)$ when $k < \sigma^2/||\tilde{\beta}||^2$ and because, as mentioned in the same article, when K has equal diagonal elements k, β_K has the intuitive appeal of minimizing f(b), subject to

$||b|| \leq ||\beta_k||$. R3B was the same as R3A, except the variables were not first standardized. The R3C estimate was constructed in the same manner as R3B, except that k was always chosen to be $\hat{\sigma}^2/||\tilde{\beta}_0||^2$; that is, no iterative procedure was used. As Table 6 in Appendix A indicates, all three of these estimators were approximately equal to OLS for r = .9. For r = .99, however, they were considerably better than OLS for θ = 0°, 15°, 30°, 45°, although considerably worse for the other values of θ. Their performance in the unfavorable cases seems to outweigh their performance for favorable θ.

V. FUTURE WORK

Since the angle θ of the true coefficient vector will obviously be unknown to the investigator, it would appear that he is at a loss when determining whether to use a ridge estimator for a particular regression problem. However, we also noted in our experiments that the value of R^2 (the sample multiple correlation coefficient) was directly correlated with θ: the smaller θ, the higher R^2 tended to be. This is because for a given matrix of cross-products S, we held the variance σ^2 of the error term constant for the different values of θ. Now R^2 is given by $R^2 = \beta_0'X'y/y'y = \beta_0'S\beta_0/y'y$. Heuristically, if we estimate $E(R^2)$ by $\beta'S\beta/N\sigma^2$, where β is the true coefficient vector (this formula is not, of course, rigorously true), then we can see that when $\theta = 0^o$ (when β is the eigenvector for the larger eigenvalue), this expression will be maximized; for $\theta = 90°$ it will be minimized. The estimate computed from this heuristic formula actually turned out to approximate the observed mean value of R^2 quite well.

In order to examine this relationship of R^2 more fully, we conducted an experiment using the same range of values of θ, but allowing σ^2 to vary with θ so that $E(R^2)$ remained at approximately .90. As Table 7 indicates, the ridge estimators tested (R3A, R3B, R3C) were virtually indistinguishable from OLS for r = .90. For r = .99, however, they outperformed OLS for $\theta = 0°$, 15°, 30°, and 45°; and for the remaining values of θ they were indistinguishable from OLS. Since for a given problem the sample R^2 is known to the investigator, it appears it might be possible to construct a ridge estimator that would be better than OLS, providing k were specified as the appropriate function of the sample at hand.

Other considerations support this conjecture. For example, suppose all the eigenvalues λ_i of S are equal, and suppose also that the dimension n is greater than 2. Baranchik (1967) has shown that the truncated Stein estimator β^* (James and Stein, 1961) is better than β_0 for all values of β. This estimator is defined by $\beta^* = (1 - \mu)^+ \beta_0$, where $\mu = c\hat{\sigma}^2/||\beta_0||^2$ for any constant c between 0 and 2(n-2)/(N-n+2)

(a^+ denotes max {a, 0}). By an appropriate choice of k (namely, $k = \frac{\mu}{1-\mu}$ for $\mu < 1$ and $k = 0$ otherwise), the ridge estimate $\beta_k = (S + kI)^{-1}X'y = (I + kI)^{-1}\beta_0$ will then coincide with β^* whenever $\beta^* \neq 0$, and will be equal to β_0 when $\beta^* = 0$.

These remarks are not strictly relevant to ridge estimation, since the problem of multicollinearity is characterized precisely by _unequal_ λ_i. We wish to suggest, however, that (as has been done with Stein estimators) it is probable that a rule for determining k could be defined, such that β_k would be better than OLS. Until the properties of such a ridge estimator are rigorously derived, however, we would caution the reader against using ridge analysis to estimate regression coefficients.

Appendix A

NUMERICAL RESULTS

The following tables contain the coefficient mean squared error (MSE) entries for the estimators tested. Two models were used:

MODEL 1

	Cross-products Matrix		Eigenvalues	Correlation Matrix		Eigenvalues
r = .9108	120.21	107.11	224.77	1.000	.9108	1.9108
	107.11	115.04	10.481	.9108	1.000	.0892
r = .9927	126.84	126.53	254.00	1.000	.9927	1.9927
	126.53	128.09	.93658	.9927	1.000	.0073

MODEL 2

	Cross-products Matrix		Eigenvalues	Correlation Matrix		Eigenvalues
r = .9108	120.21	160.67	364.51	1.000	.9108	1.9108
	160.67	258.85	14.542	.9108	1.000	.0892
r = .9927	126.84	189.79	413.75	1.000	.9927	1.9927
	189.79	288.20	1.2932	.9927	1.000	.0073

All cross-products matrices were formed from 100 observations on the independent variables.

Table 1 gives the component values for the true coefficient vectors β corresponding to the angles 0°, 15°,..., 90°.

In Tables 2-6, the error variance σ^2 was held constant at 1/9. $E(R^2)$ is the approximate formula mentioned in the text; $A(R^2)$ is the observed average value of R^2.

In Table 7, $E(R^2)$ was held constant (at .90) and σ was allowed to vary.

In Table 5, α% indicates that the ordinary least squares estimate rather than the ridge estimate was used when the statistic u was greater than u_α. The following values were used:

α	u_α
50	1.113
25	2.883
10	5.351
5	7.244

Table 1

TRUE COEFFICIENT VALUES

	Model 1		Model 2	
	β_1	β_2	β_1	β_2
r = .9108				
Angle				
0	.7156	.6985	.5495	.8355
15	.8720	.4895	.7470	.6648
30	.9690	.2472	.8936	.4488
45	.9999	-.0121	.9793	.2023
60	.9627	-.2704	.9983	-.0581
75	.8599	-.5104	.9493	-.3145
90	.6985	-.7156	.8355	-.5495
r = .9927				
Angle				
0	.7054	.7089	.5517	.8340
15	.8648	.5021	.7488	.6628
30	.9653	.2612	.8948	.4464
45	1.000	.0025	.9799	.1996
60	.9666	-.2564	.9982	-.0608
75	.8673	-.4979	.9484	-.3171
90	.7089	-.7054	.8340	-.5517

Table 2

MSE FOR 1000 TRIALS USING MODEL 1

	OLS		R1A		R1B		R1C			
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	$E(R^2)$	$A(R^2)$
r = .9108										
Angle										
0	.0054	.0056	.0024	.0022	.0023	.0021	.0039	.0040	.9529	.9537
15	.0055	.0057	.0252	.0073	.0214	.0062	.0065	.0067	.9498	.9515
30	.0053	.0056	.1026	.0412	.0950	.0377	.0060	.0063	.9391	.9411
45	.0052	.0057	.2342	.1122	.2236	.1073	.0055	.0060	.9137	.9164
60	.0050	.0051	.3514	.2101	.3194	.2016	.0052	.0053	.8522	.8566
75	.0056	.0059	.1062	.0892	.1104	.0951	.0059	.0060	.6909	.6964
90	.0053	.0054	.0053	.0054	.0053	.0054	.0053	.0055	.4854	.4963
r = .9927										
Angle										
0	.0624	.0615	.0010	.0012	.0012	.0011	.0143	.0143	.9581	.9596
15	.0587	.0576	.0353	.0214	.0364	.0223	.0357	.0353	.9552	.9566
30	.0613	.0606	.1266	.0909	.1286	.0925	.0849	.0835	.9450	.9468
45	.0553	.0547	.2550	.1952	.2573	.1971	.1158	.1140	.9198	.9218
60	.0562	.0561	.3867	.3108	.3888	.3125	.1209	.1206	.8525	.8578
75	.0622	.0615	.4906	.4092	.4880	.4113	.1250	.1233	.6169	.6273
90	.0610	.0608	.0948	.0944	.0914	.0909	.1246	.1231	.0778	.0947

Table 3

MSE FOR 100 TRIALS USING MODEL 1

	OLS		R1A		R1B		R1C			
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	$E(R^2)$	$A(R^2)$
r = .9108										
Angle										
0	.0044	.0050	.0020	.0020	.0020	.0019	.0031	.0036	.9529	.9548
15	.0043	.0044	.0252	.0068	.0212	.0057	.0052	.0052	.9498	.9509
30	.0050	.0057	.1022	.0429	.0945	.0394	.0060	.0068	.9391	.9407
45	.0056	.0062	.2339	.1131	.2233	.1082	.0060	.0067	.9137	.9158
60	.0047	.0049	.3529	.2096	.3206	.2009	.0049	.0050	.8522	.8591
75	.0060	.0067	.1298	.1107	.1374	.1203	.0063	.0069	.6909	.7013
90	.0049	.0043	.0049	.0043	.0049	.0043	.0049	.0045	.4854	.4968
r = .9927										
Angle										
0	.0806	.0768	.0014	.0013	.0015	.0012	.0253	.0232	.9581	.9595
15	.0655	.0622	.0362	.0207	.0374	.0216	.0336	.0318	.9552	.9570
30	.0638	.0637	.1252	.0937	.1270	.0953	.0873	.0872	.9450	.9476
45	.0632	.0649	.2551	.1972	.2574	.1991	.1267	.1258	.9198	.9230
60	.0485	.0479	.3880	.3110	.3900	.3127	.1125	.1111	.8525	.8576
75	.0794	.0790	.4956	.4071	.4925	.4089	.1423	.1395	.6169	.6187
90	.0681	.0645	.0887	.0839	.0757	.0717	.1080	.1053	.0777	.1073

Table 4

MSE FOR 100 TRIALS USING MODEL 2

	OLS		R1A		R1B		R1C			
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	$E(R^2)$	$A(R^2)$
r = .9108										
Angle										
0	.0054	.0023	.0212	.0290	.0021	.0011	.0060	.0006	.9704	.9710
15	.0056	.0026	.0032	.0031	.0225	.0026	.0049	.0023	.9685	.9694
30	.0065	.0031	.0087	.0018	.1124	.0182	.0060	.0027	.9614	.9629
45	.0047	.0022	.0867	.0144	.2547	.0526	.0052	.0024	.9446	.9463
60	.0046	.0022	.2816	.0619	.3534	.1008	.0051	.0024	.9018	.9031
75	.0056	.0027	.2244	.0764	.0270	.0110	.0056	.0027	.7737	.7808
90	.0064	.0032	.0064	.0032	.0064	.0032	.0067	.0033	.5669	.5756
r = .9927										
Angle										
0	.0634	.0278	.0870	.0534	.0008	.0008	.0776	.0318	.9738	.9748
15	.0589	.0264	.0093	.0082	.0454	.0127	.0268	.0125	.9720	.9735
30	.0592	.0256	.0162	.0031	.1666	.0829	.0227	.0093	.9655	.9666
45	.0603	.0253	.1176	.0357	.3356	.1132	.0712	.0300	.9492	.9510
60	.0658	.0293	.3020	.1068	.5137	.1864	.1341	.0598	.9038	.9066
75	.0813	.0359	.5324	.1947	.6388	.2456	.1505	.0662	.7225	.7325
90	.0572	.0248	.0824	.0361	.0716	.0311	.1121	.0492	.1043	.1187

Table 5

MSE FOR 100 TRIALS USING MODEL 1

	OLS		R2-50%		R2-25%		R2-10%		R2-5%			
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	$E(R^2)$	$A(R^2)$
r = .9108												
Angle												
0	.0054	.0052	.0053	.0050	.0046	.0043	.0034	.0032	.0029	.0022	.953	.954
15	.0056	.0059	.0057	.0058	.0067	.0089	.0086	.0064	.0104	.0071	.950	.951
30	.0065	.0069	.0065	.0069	.0065	.0069	.0065	.0069	.0071	.0071	.939	.941
45	.0047	.0049	.0047	.0049	.0047	.0049	.0047	.0049	.0047	.0049	.914	.916
60	.0046	.0050	.0046	.0050	.0043	.0047	.0043	.0047	.0043	.0047	.852	.854
75	.0056	.0061	.0052	.0055	.0065	.0070	.0065	.0070	.0065	.0070	.691	.701
90	.0064	.0072	.0064	.0072	.0064	.0072	.0064	.0072	.0064	.0072	.485	.494
r = .9927												
Angle												
0	.0632	.0625	.0626	.0618	.0555	.0547	.0334	.0336	.0266	.0272	.958	.960
15	.0588	.0594	.0590	.0581	.0618	.0567	.0599	.0513	.0565	.0459	.955	.958
30	.0592	.0575	.0609	.0578	.0759	.0676	.0915	.0761	.1132	.0914	.945	.947
45	.0603	.0569	.0623	.0579	.0649	.0598	.0934	.0809	.1259	.1054	.920	.923
60	.0658	.0659	.0657	.0653	.0789	.0752	.1205	.1097	.1517	.1357	.852	.857
75	.0812	.0806	.0808	.0796	.0894	.0867	.1423	.1307	.1585	.1442	.617	.630
90	.0572	.0558	.0619	.0604	.0827	.0812	.0827	.0812	.0827	.0812	.078	.093

Table 6

MSE FOR 100 TRIALS USING MODEL 2

	OLS β_1	OLS β_2	R3A β_1	R3A β_2	R3B β_1	R3B β_2	R3C β_1	R3C β_2	$E(R^2)$	$A(R^2)$
r = .9108										
Angle										
0	.0054	.0023	.0054	.0023	.0054	.0023	.0053	.0023	.9704	.9710
15	.0056	.0026	.0056	.0026	.0056	.0026	.0055	.0026	.9685	.9694
30	.0065	.0031	.0065	.0031	.0065	.0031	.0065	.0031	.9614	.9629
45	.0047	.0022	.0047	.0022	.0047	.0022	.0047	.0022	.9446	.9463
60	.0046	.0022	.0046	.0022	.0046	.0022	.0048	.0023	.9018	.9031
75	.0056	.0027	.0057	.0028	.0056	.0027	.0055	.0027	.7737	.7808
90	.0064	.0032	.0067	.0034	.0067	.0033	.0066	.0033	.5669	.5756
r = .9927										
Angle										
0	.0633	.0278	.0578	.0254	.0600	.0263	.0555	.0244	.9738	.9748
15	.0589	.0264	.0557	.0250	.0533	.0240	.0523	.0235	.9720	.9735
30	.0592	.0256	.0539	.0232	.0539	.0231	.0550	.0236	.9655	.9666
45	.0603	.0253	.0513	.0214	.0547	.0228	.0555	.0232	.9492	.9510
60	.0658	.0293	.0659	.0294	.0712	.0318	.0709	.0317	.9038	.9066
75	.0813	.0359	.0951	.0418	.0970	.0427	.0896	.0394	.7225	.7325
90	.0572	.0248	.1150	.0507	.1150	.0504	.0706	.0308	.1043	.1187

Table 7

MSE FOR 100 TRIALS USING MODEL 2

	OLS		R3A		R3B		R3C			
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	σ	$A(R^2)$
r = .9108										
Angle										
0	.0197	.0084	.0197	.0084	.0197	.0084	.0188	.0080	.6364	.9022
15	.0192	.0090	.0192	.0090	.0192	.0090	.0184	.0086	.6156	.9031
30	.0179	.0085	.0179	.0085	.0179	.0085	.0180	.0084	.5548	.9034
45	.0088	.0042	.0088	.0042	.0088	.0042	.0089	.0042	.4589	.9029
60	.0047	.0023	.0047	.0023	.0047	.0023	.0049	.0023	.3367	.9013
75	.0021	.0010	.0021	.0010	.0021	.0010	.0021	.0010	.2054	.9031
90	.0009	.0005	.0009	.0005	.0009	.0005	.0009	.0005	.1271	.9038
r = .9927										
Angle										
0	.2618	.1150	.2043	.0898	.1871	.0825	.1860	.0820	.6780	.9036
15	.2273	.1021	.1685	.0762	.1652	.0747	.1657	.0749	.6550	.9050
30	.1840	.0794	.1362	.0584	.1454	.0622	.1482	.0633	.5875	.9028
45	.1252	.0525	.0967	.0402	.1085	.0451	.1080	.0449	.4802	.9035
60	.0687	.0306	.0686	.0306	.0747	.0334	.0741	.0331	.3406	.9029
75	.0235	.0104	.0247	.0109	.0245	.0108	.0246	.0108	.1793	.9043
90	.0007	.0003	.0007	.0003	.0007	.0003	.0007	.0003	.0379	.9031

Appendix B

GEOMETRIC PROPERTIES

We shall consider here some of the geometric properties of ridge estimators. These elementary properties have been mentioned and to varying degrees proved in Hoerl (1962) and Hoerl and Kennard (1970a); we present proofs here for the sake of completeness.

Only estimators β_K where K = k will be discussed, since these geometric considerations do not apply with arbitrary diagonal K. The reader should recall the definition of the orthogonal model (2.3); it is particularly useful when visualizing these geometric properties. Recall that

$$\beta_k = (S + k)^{-1}X'y$$

and

$$\tilde{\beta}_k = (D + k)^{-1}\tilde{X}'y. \qquad (B.1)$$

We shall regularly denote X'y by c and $\tilde{X}'y$ by $\tilde{c}$.

If we view the residual error sum of squares as a function of the estimate $(f(b) = (Xb - y)'(Xb - y))$, we know that the OLS estimate β_0 minimizes f globally; we shall show that the ridge estimate yields a constrained minimum for f.

<u>Theorem 1</u>: Let $||\beta_k|| = r$, where $k > 0$. Then β_k is the unique vector that minimizes f(b) subject to $||b|| \leq r$.

The following lemma shows that $r < ||\beta_0||$; thus a ridge estimate is the vector that minimizes f(b), subject to the constraint that the length of b be no greater than a value less than the length of the OLS estimate β_0. If $r \geq ||\beta_0||$ is allowed, then the constrained minimum is of course just the global minimum β_0.

We shall denote $\{b: ||b|| \leq r\}$ by B_r. In order to prove this theorem we need several lemmas.

<u>Lemma 1:</u> Suppose $0 \leq s < t$. Then

$$||\beta_t|| < ||\beta_s|| \quad . \qquad (B.2)$$

Proof: Since T preserves lengths, (B.2) is equivalent to $||\tilde{\beta}_t|| < ||\tilde{\beta}_s||$. But $||\tilde{\beta}_t||^2 = \tilde{\beta}_t{}'\tilde{\beta}_t = \sum_i \frac{\tilde{c}_i^{\,2}}{(\lambda_i + t)^2}$, and similarly for $\tilde{\beta}_s$; and since $(\lambda_i + t)^2 > (\lambda_i + s)^2$ for all i, this completes the proof.

We next show that in seeking our minimizing vector it suffices to consider the surface of the sphere B_r.

Lemma 2: Let $r < ||\beta_0||$, and suppose β^* minimizes f(b) subject to $||b|| \leq r$. Then

$$||\beta^*|| = r \ . \qquad (B.3)$$

Proof: Suppose $||\beta^*|| < r$. Since β^* is thus strictly inside B_r, there is a small neighborhood about β^* such that β^* minimizes f on this neighborhood; that is, β^* is a (unconstrained) local minimum for f. But then the gradient of f, ∇f, must satisfy $\nabla f(\beta^*) = 0$ (∇f is the vector-valued function whose ith component is given by $\frac{\partial f}{\partial b_i}$). It is easily seen that $\nabla f(b) = 2Sb - 2c$, so β^* must satisfy $S\beta^* = c$. Since the only solution to this is β_0, we have $\beta^* = \beta_0 \notin B_r$, a contradiction; hence $||\beta^*|| < r$ cannot hold.

We now determine a necessary condition for a vector to yield a constrained minimum for f.

Lemma 3: Suppose β^* minimizes f(b) subject to $||b|| = r$. Then there exists a real number λ such that

$$(S + \lambda)\beta^* = c \ . \qquad (B.4)$$

Proof: Let $g(b) = ||b||^2$; we wish to minimize f(b) subject to $g(b) = r^2$. By a theorem due to Lagrange, β^* must satisfy $\nabla f(\beta^*) + \lambda \nabla g(\beta^*) = 0$ for some real λ. Thus $2S\beta^* - 2c + 2\lambda\beta^* = 0$, which is equivalent to (B.4).

There may, of course, be more than one λ such that there exists a β^* satisfying (B.4). The next lemma indicates which λ to consider when seeking a constrained minimum.

<u>Lemma 4</u>: Suppose $\lambda < \mu$, and suppose a and b are two distinct vectors such that

$$\text{(i)}\quad ||a|| = ||b|| = r > 0,$$

$$\text{(ii)}\quad (S + \lambda)a = c,$$

and

$$\text{(iii)}\quad (S + \mu)b = c.$$

Then

$$f(b) < f(a). \tag{B.5}$$

<u>Proof</u>: From (ii) and (i) we obtain $a'Sa = a'c - \lambda a'a = a'c - \lambda r^2$, so that $f(a) = a'Sa - 2a'c + y'y = y'y - a'c - \lambda r^2$. Writing f(b) in a similar manner reduces the proof of (B.5) to showing

$$a'c - b'c < (\mu - \lambda)r^2 . \tag{B.6}$$

The left hand side of this inequality may be written as

$$a'(S + \mu)b - b'(S + \lambda)a = (\mu - \lambda)a'b .$$

By Schwarz's inequality $|a'b| \leq ||a||\ ||b|| = r^2$, with equality holding if and only if $a = tb$ for some real t. If strict inequality holds, we see that $\mu - \lambda > 0$ implies $(\mu - \lambda)a'b < (\mu - \lambda)r^2$, proving (B.6). If $a = tb$, then by (i) $|t| = 1$. The assumed distinctness of a and b then implies $t = -1$, so that $(\mu - \lambda)a'b = -(\mu - \lambda)r^2$, again resulting in (B.6).

We may now prove the theorem.

<u>Proof of Theorem 1</u>: We first observe that f does attain a minimum on B_r, since f is continuous and the sphere is compact. Let β^* be a point at which this constrained minimum is attained; we need to show $\beta^* = \beta_k$. As remarked earlier, $r < ||\beta_0||$; thus by Lemma 2, $||\beta^*|| = r$. Lemma 3 then gives the existence of a real λ such that

$$(S + \lambda)\beta^* = c . \tag{B.7}$$

$k < \lambda$ is impossible, since Lemma 1 would then imply $||\beta^*|| < ||\beta_k|| = r$. $\lambda < k$ cannot hold either, for then $||\beta^*|| = ||\beta_k||$ together with Lemma 4 would result in $f(\beta_k) < f(\beta^*)$, contradicting β^*'s minimality (unless, of course, $\beta^* = \beta_k$). Hence $\lambda = k$, so β_k also satisfies (B.7). Since $\lambda > 0$ implies $S + \lambda$ is non-singular, this proves $\beta^* = \beta_k$.

This shows, incidentally, why β_k is referred to as a "ridge" estimate. For if one views the graph of f(b) as a subset of $\mathbb{R}^{n+1}$, the portion $R = \{(\beta_k, f(\beta_k)): k > 0\}$ defines the lower ridge of the graph as one moves from $(\beta_0, f(\beta_0))$ in toward the origin. Thus if $n = 2$, for a given value of r the corresponding point $(\beta_k, f(\beta_k))$ will be the "lowest" point that can be obtained on the graph when $||\beta|| = r$.

Another way of stating Theorem 1 is the following:

<u>Theorem 2</u>: Let

$$\text{(i)} \quad f(\beta_0) < \gamma < y'y \ .$$

Then there exists a $k > 0$ such that β_k is the unique vector that minimizes $||b||$, subject to $f(b) \leq \gamma$.

We thus see that a ridge estimate is the shortest estimate possible, subject to the constraint that the residual sum of squares not increase (above the minimum residual at β_0) by more than a specified amount.

<u>Proof</u>: Define the function φ on $[0,\infty)$ by $\varphi(\lambda) = f(\beta_\lambda)$; clearly φ is continuous. $\varphi(0) = f(\beta_0)$ and $\varphi(\lambda) \to y'y$ as $\lambda \to \infty$, so by (i) there must exist a $k > 0$ such that $\varphi(k) = \gamma$. Suppose $\beta^* \neq \beta_k$ and $||\beta^*|| \leq ||\beta_k|| = r$. By Theorem 1, since β_k is the unique point minimizing f(b) subject to $||b|| \leq r$, we must have $f(\beta^*) > f(\beta_k) = \gamma$. There is thus no other vector β^* as short as β_k such that $f(\beta^*) \leq \gamma$, proving the theorem.

Geometrically, it can be seen that the set $E = \{\beta: f(\beta) \leq \gamma\}$ is an n-dimensional (solid) ellipsoid, centered at β_0, and having principal axes corresponding to the eigenvectors of S. Theorem 2 then states that the ridge estimate is the point on (the surface of) this ellipsoid that is closest to the origin.

REFERENCES

Baranchik, A. J. (1967), "A Family of Minimax Estimators of the Mean of a Multivariate Normal Distribution," Technical Report, Department of Mathematical Statistics, Columbia University, New York.

Hoerl, A. E. (1959), "Optimum Solution of Many Variables Equations," Chemical Engineering Progress 55, pp. 69-78.

----- (1962), "Application of Ridge Analysis to Regression Problems," Chemical Engineering Progress 58, pp. 54-59.

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

----- (1970b), "Ridge Regression: Applications to Nonorthogonal Problems," Technometrics 12, pp. 69-82.

James, W. and C. Stein (1961), "Estimation with Quadratic Loss," Proceedings of the Fourth Berkeley Symposium, Vol. 1, University of California Press, pp. 361-379.

Toro-Vizcarrondo, C. and T. D. Wallace (1968), "A Test of the Mean Square Error Criterion for Restrictions in Linear Regression," Journal of the American Statistical Association 63, pp. 558-572.

Wallace, T. D. and C. Toro-Vizcarrondo (1969), "Tables for the Mean Square Error Test for Exact Linear Restrictions in Regression," Journal of the American Statistical Association 64, pp. 1649-1663.