This is to certify that the

thesis entitled

A MONTE CARLO EVALUATION OF RIDGE REGRESSION
AS AN ALTERNATIVE TO ORDINARY LEAST SQUARES

presented by

BRYAN WALTER COYLE

has been accepted towards fulfillment
of the requirements for

M.A. degree in PSYCHOLOGY

Neal Schmitt
Major professor

Date 11-1-79

A MONTE CARLO EVALUATION OF RIDGE REGRESSION
AS AN ALTERNATIVE TO ORDINARY LEAST SQUARES

By

Bryan Walter Coyle

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF ARTS

Department of Psychology

1979

ABSTRACT

A MONTE CARLO EVALUATION OF RIDGE REGRESSION AS AN ALTERNATIVE TO ORDINARY LEAST SQUARES

By

Bryan Walter Coyle

This study investigated a proposed modification of ordinary least squares (OLS) multiple regression. Conventional OLS is generally used to combine the information present among a set of variables so as to optimize the prediction of a criterion variable in the original sample and to provide an equation for use in subsequent samples without the necessity of re-estimation. In addition, the predictor weights estimated are frequently used to infer the functional characteristics of the system which produced the data.

Hoerl and Kennard (1970a) have suggested deliberately introducing a statistical bias into the OLS estimation procedure in an attempt to increase the predictive robustness and structural accuracy of ordinary least squares in collinear data sets. Their method, termed ridge regression, was compared with unit weighting (Schmidt, 1971) and with OLS in a Monte Carlo experiment based on three data matrices drawn from the literature.

It was concluded that the ridge technique can outperform OLS in situations where the collinearity is high and consistent across all predictors. When the collinearity is concentrated in subsets of the predictor matrix ridge regression is dominated by OLS. Consistent with Schmidt (1971), when sample sizes are small relative to the number of predictors, no suppressors are present and only prediction as opposed to structural interpretation is relevant, unit weighting is to be preferred.

Approved:______________________________

Date:__________________________________

Thesis Committee:

Neal Schmitt, Chairperson

Raymond Frankmann

Ralph Levine

ACKNOWLEDGMENTS

I wish to thank Dr. Neal Schmitt, my chairperson, for both his assistance on this and many other projects and for his sage counsel and friendship over several years.

Thanks are also due to Dr. Raymond Frankmann and Dr. Ralph Levine for their advice and guidance as committee members and teachers.

Finally, I thank my wife, Phyllis, without whose love and support this and many other endeavors would have little meaning.

TABLE OF CONTENTS

Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Linear composites are commonly used in psychology, education and in the social sciences generally for the purpose of combining the information present among a group of variables into a single variable. While many methods of forming a composite have been proposed and utilized (Blum & Naylor, 1968; Burket, 1964; Claudy, 1972; Lawshe & Shucker, 1959) multiple linear regression is by far the most commonly used combinatorial method. This is especially true since the advent of digital computers and widely available "canned" programs which save the researcher the tedium of hand calculation and often the concomitant necessity of considering the applicability of this method to the research problem at hand.

Linear composites, whether formed by multiple regression or other techniques to be discussed subsequently, are generally used for either predictive or descriptive purposes. In the former case one is interested in creating the composite so as to maximize its correlation with some external variable, usually designated the criterion. Examples of this usage would be predicting a person's future academic standing from past records or estimating the probability of job success on the basis of a composite formed by qualification tests, interview data and previous employment history. Descriptive uses, also termed structural interpretation, of linear composites involve assessing the degree of change produced in the criterion variable by a unit change in one or more of those variables which form the composite.

As multiple linear regression is the most frequently used combinatorial scheme, at least when the number of available cases is large relative to the number of indicator variables intended to form the composite, its assumptions and use will be discussed first. Limitations inherent in this model are presented as well as the major alternatives to it. Various criteria that have been proposed to evaluate the optimality of combination rules are then contrasted with a modified regression approach, termed ridge regression (Hoerl & Kennard, 1970a; 1970b). The empirical performance of this method was assessed in a Monte Carlo design employing three data sets with different degrees of intervariable relationships. From each of these populations 25 samples at each of five sample sizes were randomly drawn. Ridge regression and ordinary least squares (multiple regression) are then employed on each of these 375 samples as are three different methods of simple unit weighting (Schmidt, 1971). For each method the predictive efficiency in the initial sample as well as the long-term efficiency in the population are evaluated. In addition, the structural accuracy with respect to the precision of parameter estimation of ridge regression and ordinary least squares (OLS) are compared.

CHAPTER II

MULTIPLE LINEAR REGRESSION

The Model

An optimal method for obtaining estimates of criterion values as a function of predictor score levels would be the following (Burket, 1964): Select all conceptually relevant variables not statistically independent of the criterion. Measurements on these predictors and the criterion should be obtained on a sufficiently large number of cases (termed the validation sample) such that all possible combinations of score levels are represented. The criterion prediction for a particular case would be the criterion mean of all cases in the validation sample having the same predictor profile.

In practice this idealized system is not generally workable because of the large sample size required to insure stable parameter estimates for every possible predictor profile. What is necessary then is to make simplifying assumptions and adopt a system which will provide fairly accurate predictions of criterion performance over a wide range of possible predictor profiles despite the unavailability of some of them. The assumption most frequently employed in the behavioral sciences is that there exists an approximate functional relationship (most often presumed to be linear although this is not necessary) between the predictors and criterion. The function form relating these is estimated in the sample at hand by the method of multiple linear regression or ordinary least squares which assures two important properties: (1) the sum of squared residuals between the actual criterion values and those predicted from the weighted profile components will be minimized for the validation sample; and (2) the correlation between these two score sets will be the maximum obtainable for this sample (Draper & Smith, 1966; Li, 1974).

The method of least squares and its properties may be summarized with the following notation. The linear regression function relating the dependent variable (Y) to one or more independent predictor variables (X), is for the i^{th} case,

$$\hat{y}_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + . . . + \beta_p x_{ip} + \varepsilon_i \tag{1}$$

where: y_i = criterion score for the i^{th} subject in the sample,

α = a scaling constant used to adjust for differences in origin between the y and x variables; also termed the intercept,

β_j = a partial regression coefficient used to weight the j^{th} predictor variable,

x_{ij} = the i^{th} individual's observed score on the j^{th} predictor,

ε_i = error in prediction for the i^{th} subject,

$\hat{y}_i$ = predicted criterion score for the i^{th} subject.

Thus,

$$(2) \qquad \varepsilon_i = y_i - (\hat{\alpha} + \sum_{j=1}^{p} \hat{\beta}_j x_{ij})$$
$$= y_i - \hat{y}_i$$

Thus the properties of OLS noted above are then

$$(3) \qquad \Sigma_i\ (y_i - \hat{y}_i)^2 = \text{minimum}$$

$$(4) \qquad r_{y\hat{y}} = \text{maximum},$$

with these properties holding for the N cases in the sample on which the weights were estimated. The correlation of equation (4) is referred to as the multiple correlation or if squared, the coefficient of determination of the weighted predictor composite with the criterion.

The model for estimating the weights is more easily presented in matrix terms and that notation will be established here. Proofs of the derivations are available in Draper and Smith (1966), Finn (1974) and Scheffe (1959). Without loss of generality the observations on all variables are assumed to be standardized so that the constant term (α) in the general model is identically zero.

Let y be a column vector of N criterion observations,

X be a N x p matrix with rank p less than N, each row representing one cases' observations on the p predictor variables,

ε be a column vector of N uncorrelated errors with mean zero and variance σ^2,

β be a column vector of p population regression coefficients.

The general linear model presented in (1) becomes

(5) $$y = X\beta + \varepsilon.$$

Because of the assumptions concerning errors, $E(\varepsilon) = 0$; $E(\varepsilon\varepsilon') = \sigma^2 I$, the criterion vector y has the expectation

(6) $$E(y) = X\beta$$

and the covariance matrix

(7) $$E[(y - X\beta)\ (y - X\beta)'] = \sigma^2 I.$$

If the $\hat{\beta}$ are the sample estimates of the population regression coefficients, β, and $\hat{y}$ are the predicted criterion scores based on these same sample estimates then

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

and,

(9) $$\hat{y} = X\hat{\beta}.$$

Because the variables have been standardized (X'X) is in the form of a zero-order correlation matrix among the predictors and X'y is the vector of predictor-criterion correlations or validities. The estimates of the population regression coefficients have the expectation

(10) $$E(\hat{\beta}) = \beta$$

and the covariance matrix

(11) $$E\ [(\hat{\beta} - \beta)\ (\hat{\beta} - \beta)'] = \sigma^2(X'X)^{-1}.$$

$\hat{\beta}$ is the "best" estimate of the population vector β in that the sum of squared errors in prediction is minimized in the sample. This can be demonstrated (Finn, 1974, p. 96) by considering any other estimate β^* where $\beta^* = \hat{\beta} + d$ and d is the vector of discrepancies between β and the alternative estimate. The sum of squared errors with β^* replacing $\hat{\beta}$ is

$$\begin{aligned}(\varepsilon'\varepsilon)^* &= (y - X\beta^*)'(y - X\beta^*)\\ &= [(y - X\hat{\beta}) - Xd]'[(y - X\hat{\beta}) - Xd]\\ &= (y - X\hat{\beta})'(y - X\hat{\beta}) - 2d'X'(y - X\hat{\beta}) + d'X'Xd.\end{aligned}$$

The first term is $\varepsilon'\varepsilon$; the second term is zero as from equation (8)

$$d'X'(y - X\hat{\beta}) = d'X'y - d'X'X(X'X)^{-1}X'y = 0.$$

The third term is positive as it represents the sum of the squared elements of Xd. Thus the residuals $(\varepsilon'\varepsilon)$ are inflated anytime one departs from $\hat{\beta}$ as the estimate of population weights derived from the sample.

The variance of these minimized residuals in the standardized case is one minus the coefficient of multiple determination (R^2), that value which expresses the squared correlation between the optimally weighted predictor combination and the criterion. Where X'X is in the form of a correlation matrix equivalent formulae for R^2 are (Burket, 1964; Overall & Klett, 1972),

(12) $$R^2 = y'X(X'X)^{-1}X'y$$

(13) $$R^2 = \hat{\beta}'X'y.$$

R^2, R or $(1 - R^2)$ are commonly presented as indices of the predictive efficiency of the multiple regression model in the estimation sample.

Although the multiple linear regression model presented above has been used extensively throughout the sciences for the purposes of prediction and structural interpretation, its assumptions are often poorly understood. Cureton (1950) considered that, "It is doubtful that any other statistical techniques have been so generally and widely misused and misinterpreted as have those of multiple correlation" (p. 690). The situation is perhaps worse today with the wider availability of canned computer programs for regression.

Assumptions of the Regression Model

The simplest set of crucial assumptions (Johnston, 1972, p. 122) necessary to estimate the β vector in the model $y = X\beta + \varepsilon$ are three in number:

(14) $E(\varepsilon\varepsilon') = \sigma^2 I$,

(15) X is a set of fixed values,

(16) X has rank p less than N.

The requirement of (14) is that the error or disturbance values have constant variance--a property referred to as homoscedasticity. The diagonal nature of the symmetric matrix $E(\varepsilon\varepsilon')$ implies that the covariance between any pair of error terms be zero. Fulfillment of this assumption can often be evaluated by visual examination of residual value plots (Draper & Smith, 1966, ch. 3). Failure to meet this assumption most often occurs in time series analysis or when the linear model fitted is inappropriate for the set of observations at hand (i.e., there exists a nonlinear relation between the predictors and the criterion). As the assumption can generally be adequately met by the inclusion of appropriate quadratic terms or by the inclusion of linear or higher order terms in time or by suitable data transformations such as the arcsin, square root or log transforms before analysis (Tukey, 1949; Winer, 1971) the consequences of failure to meet this assumption will not be considered further here.

The second essential assumption (15) is more germane to the purpose of the present paper. Regression theory requires that the X matrix be a set of values fixed by the experimenter exactly as are the levels of independent variables at which observations on y, the criterion, are taken in fixed effects analysis of variance designs (Binder, 1959). Implicit in this assumption is the requirement that the X values be free of measurement error. This means that in repeated sampling of criterion values the only source of variation is attributable to the vector of disturbances, ε. If this assumption is met, $\hat{\beta}$ is an unbiased linear estimator (Johnston, 1972, pp. 18-23). Effects of violations of this assumption are discussed below under the correlation model.

The third assumption (16) states that X must be of full rank equal to p, the number of predictors. If the rank of X is less than the number of predictors the $\hat{\beta}$ vector is indeterminate and no unique solution to the normal equations exists. As will be discussed under the heading of "multicollinearity," problems can also arise when this assumption is only approximately met.

The Correlation Model

While data transformations or deletion of some predictors have been found in many cases to adequately compensate for violations of assumptions (14) and (16), failure to obtain fixed predictor values requires an alternative model. Traditionally, multiple regression techniques have been applied in precisely those situations where the control required to obtain fixed-X values cannot be insured (Cohen, 1968). Data sets analyzed by means of OLS are typified by subjects' test scores, historical records and in general by data that is not collected according to a design for the systematic evaluation of criterion scores obtained at preselected levels of the independent variables. In this type of situation the correlational model for the predictors is more appropriate than is the regression model. The latter is based on the assumption that only the disturbance vector ε is subject to sampling error--an assumption that is rarely met in applied multiple regression situations. The correlation or random-X model assumes that the predictors and the criterion are random variables sampled from a joint multivariate normal distribution.

Regardless of the distributional form of the disturbances (and hence of the y values) the OLS method provides "best"--i.e., minimum variance, unbiased estimators of the population β values. While the fixed-X or regression model makes no assumptions about the distribution of the predictor variables it does require a normal error assumption to permit inferential tests. This assumption is based on empirical evaluations of the robustness of the _t_ and _F_ statistics against moderate departures from normality (Neter & Wasserman, 1974). When this assumption is met the $\hat{\beta}$ estimates are maximum likelihood estimates of the true, population weights with the same best linear unbiased properties as the least squares values (Herzberg, 1969; Neter & Wasserman, 1974).

While both models would appear to provide the necessary data for inferential uses of multiple regression results it is clear that the correlational model is almost always more appropriate. Under the null hypothesis of zero multiple correlation the distributional theory is identical for the two models. However in applications, especially those for predictive purposes such as in personnel selection, the null hypothesis is rarely true (Burket, 1964). The extreme complexity of the correlational model in cases where the null hypothesis does not obtain has led most investigators to use the fixed-X model in the hope that there will be little practical difference in the results derived (Burket, 1964; Claudy, 1972; Cohen & Cohen, 1975; Neter & Wasserman, 1974).

While this subject has not received a great deal of attention in the literature, it would appear that application of fixed-X procedures to random-X data affects the suitability of the weight estimates thus derived. It has been demonstrated by Berkson (1950), Geary (1953) and Rao and Miller (1971) that if the predictor variables are not held at preselected values and are subject to errors of measurement the beta weights will in fact be biased estimates of the population values. It is thus not necessarily the case that beta weights derived on a sample of finite size such that they maximize the multiple correlation in the sample are the "best" estimates of the parameter vector β (Claudy, 1972). "Best" here refers not to the minimum variance properties of least squares estimators but rather to the minimization of the difference between the true population multiple correlation (ρ) and the estimate of it (R) obtained by application of the sample weights to the population.

Application of the fixed-X regression model to data collected under the assumptions of the random-X model results in an over-fitting of the regression surface to the sample data. In practice this means that the beta weights are optimized on the idiosyncracies caused by sampling and measurement error in the estimation sample. Accordingly, when these weights are applied in a new sample or in the population the resulting multiple correlation will be lower than the initial estimate. This general problem has been termed "shrinkage" and has been considered by numerous authors. For instance, formulae have been evaluated to estimate this shrinkage (see Schmitt, Coyle, & Rauschenberger, 1977 for a comparison of the major formulae) and alternatives to OLS have been proposed (Claudy, 1972; Cureton, 1962; Herzberg, 1969; Lawshe & Schucker, 1959; Schmidt, 1971).

While accurate estimation of the multiple correlation is of primary interest for predictive uses of multiple regression techniques, it does not touch upon the second purpose of multiple regression: structural interpretation.

The final consideration relative to assumption (15) is germane to this purpose. Application of fixed-X OLS to random-X data subject to sampling and measurement error inflates the variability among the optimizing weights without regard to the true variance of the parameter values. It is not the standard error or variance of any single weight estimate which is referred to here but rather the dispersion of the p weights calculated on the p predictors in a sample. This effect is attributable to the sample values being subject to not only the variance of the population weights but also to the error variance generated by the less than perfectly reliable measurement of predictor scores. Awareness of this artifact has led to such proposals as averaging the beta weights obtained in a random split of the sample (Claudy, 1972) and using as a $\hat{\beta}$ estimate the least deviant (from zero) of the two weights obtained in a fifty-fifty split (Cureton, 1962). The relative merits of several such alternatives to OLS are discussed subsequent to the further examination of the implications of assumption (16) in the following section.

Multicollinearity

The assumption of equation (16) that X, the predictor matrix, has rank $p < N$ actually has implicit two requirements. The first concerns the ratio of the sample size available to the number of predictors for which weights must be estimated, and the second involves the number of linear dependencies among the predictor set.

The question of sample size is common to any attempt to establish a statistical estimate of a parameter--the greater the number of cases upon which the estimate is derived, the more stable it will be. In the case of OLS using standardized data the requirement is that $p \leq N$ or else the model is considered to be overdefined and a unique solution for $\hat{\beta}$ is not possible. In point of fact one generally desires that N be much greater than the number of predictors for as was demonstrated by Wishart (1931),

$$(17) \qquad E(R^2) = \rho^2 + \frac{p(1 - \rho^2)}{N}$$

where ρ represents the population multiple correlation. In the case where the null hypothesis of no predictor-criterion correlation holds in the population, equation (17) shows that the sample value will be inflated. Setting ρ^2 to zero yields

$$(18) \qquad E(R^2) = \frac{p}{N}.$$

It is from this equation that the various shrinkage estimators have been derived (Darlington, 1968; Lord, 1950; Nicholson, 1960; Wherry, 1931). From the above formula it is obvious that the extent to which R^2 over-estimates ρ^2 varies directly with the number of predictors and inversely with the sample size. These characteristics are important when one compares the efficiency of OLS and alternative estimators (Schmidt, 1971) in a variety of practical situations.

Throughout the long history of multiple regression usage in the sciences, practitioners have come to appreciate its robustness in the face of violations of some underlying assumptions and have, in many cases, developed remedial procedures to correct unsuitable data before analysis. Examples of this would be the Durbin-Watson (1950) test statistic for autocorrelated error terms with the attendant suggestions of Cochrane and Orcutt (1949) as to their correction. Similarly, Bartlett's variance homogeneity test has given rise to a number of data transformations suitable to different types of heteroscedasticity (Winer, 1971). The development of detection and correction methods for problems of multicollinearity in regression models has not yet reached the level of rote application of specified test statistics which in turn could provide evidence as to appropriate alterations to be made (Farrar & Glauber, 1967). In fact, while economists have apparently been aware of the difficulties inherent in highly correlated predictor sets for some time, it seems that others in the social and behavioral sciences have frequently labeled such a concern as being of "theoretical interest only" and thereby dismissed it from consideration in their applied work. As shall be demonstrated multicollinearity can cause some very practical problems to arise (Darlington, 1968).

While a variety of definitions of multicollinearity exist in the literature, many of them are more symptomatic than definitive. The definition used here is attributable to Johnston (1972) and Silvey (1969). If one considers the predictor matrix X of dimensions (N x p) a linear dependence is said to exist between the column vectors x_1, $x_2, \ldots, x_p$ if there exist constants $a_1, a_2, \ldots, a_p$, not all zero, such that

$$(19) \qquad \sum_{i=1}^{p} a_i x_i = 0 \ .$$

When (19) holds for some subset of the column vectors of X (and thus for the matrix as a whole), multicollinearity is said to exist. In this case beta estimates cannot be obtained as the predictor matrix is singular and thus its inverse does not exist (equation 8). However, even when (19) does not obtain exactly but rather is only approximately true, multicollinearity is still a relevant problem for the data analyst. Thus the question of collinearity is one of severity or the degree of departure from orthogonal variates (Kmenta, 1971; Mason, Gunst & Webster, 1975).

There are three primary sources of highly collinear data sets (Mason, Gunst, & Webster, 1975). The first involves an overdefined model--one where there exist more predictors than observations. The difficulties caused by cases in which this is approximately true were discussed above. When faced with such a situation the analyst must, (a) eliminate some predictors; (b) use grouped subsets of predictors; or (c) utilize some form of principal components regression. There are deficiencies inherent in each of these solutions and they will be discussed in the following section.

The latter two sources of collinearity, sampling techniques and physical constraints on the model, are quite similar and can be presented together. These situations arise when the data have been sampled from only a subspace of the predictor variable domain or when some predictors' values are restricted to a near exact relationship with other variables in the X matrix. In the former case data observations can be added from the undersampled area of the domain, if indeed the investigator is aware of the problem, which can usually be identified through eigenvector analysis (Silvey, 1969). When practical constraints eliminate this alternative or when the problem cannot be identified as attributable to undersampling, few remedial measures are available.

The effects of approximate multicollinearity have been presented by Johnston (1972) as follows:

1. The precision of estimation falls so that it becomes very difficult, if not impossible, to disentangle the relative influences of the various x variables. This loss of precision has three aspects: specific estimates may have very large errors; these errors may be highly correlated, one with another; and the sampling variances of the coefficients will be very large.

2. Investigators are sometimes led to drop variables from an analysis because their coefficients are not significantly different from zero, but the true situation may be not that a variable has no effect but simply that the set of sample data has not enabled us to pick it up.

3. Estimates of coefficients become very sensitive to particular sets of sample data, and the addition of a few more observations can sometimes produce dramatic shifts in some of the coefficients (p. 160).

The first difficulty has been well documented and illustrated by Darlington (1968) while the latter two consequences are familiar to psychologists under the general rubric of "bouncing betas." The detection and analysis of these effects will be considered for the two predictor case although all results are applicable to the case of any number of predictors as long as equation (19) is not exactly satisfied.

The effects of collinearity on estimates can best be seen by considering the inverse of the correlation matrix. Equation (8) can be written for the standardized model with $C = (X'X)^{-1}$ as

$$(20) \qquad \begin{bmatrix} \hat{\beta}_{y1.2} \\ \hat{\beta}_{y2.1} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \begin{bmatrix} r_{y1} \\ r_{y2} \end{bmatrix}$$

where r_{yj} represents the validity coefficient for the jth predictor. From (20) it is evident that in the case of uncorrelated predictors ($c_{12} = c_{21} = 0$), the validity coefficient is the beta for any one predictor as the predictor matrix is then an identity matrix.

$$(21) \qquad \begin{bmatrix} \hat{\beta}_{y1.2} \\ \hat{\beta}_{y2.1} \end{bmatrix} = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix} \begin{bmatrix} r_{y1} \\ r_{y2} \end{bmatrix}$$

When the predictor intercorrelation is not equal to zero the inverse matrix is of the form

$$(22) \qquad C = (X'X)^{-1} = \left(\frac{1}{1 - r_{12}^2}\right) \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}$$

This is the matrix formulation of the familiar computational solution for beta weights with two predictors.

$$\hat{\beta}_1 = \frac{r_{y1} - r_{y2}\, r_{12}}{1 - r_{12}^2}$$

(23)

$$\hat{\beta}_2 = \frac{r_{y2} - r_{y1}\, r_{12}}{1 - r_{12}^2}$$

Equation (23) is a well-known result and illustrates that as (19) becomes exact (i.e., r_{12} or $r_{12}^2 \rightarrow 1.0$) the diagonal elements of the inverse matrix approach infinity ($1.0 \leq C_{ii} \rightarrow \infty$). Several consequences follow from this. The limiting case of intercorrelation is derived by assuming $r_{y1} = r_{y2}$ which is justified since as $r_{12} \rightarrow 1.0$, each predictor's validity must become equal (Klein & Nakamure, 1962; Sastry, 1970). This further implies that as $r_{12} \rightarrow 1.0$ the slightest discrepancy in the magnitude of the validity coefficients will result in the beta weights being approximately equal but opposite in sign. Obviously, the slight discrepancies causing this are sample specific and due to sampling error and lack of perfect reliability in measurement. This is the "bouncing beta" problem (McDonald & Schwing, 1972; Swindel, 1974; Wampler, 1970) demonstrated by sign reversals and in the case of multiple predictors by dramatic shifts in the magnitude of weights in different samples from the same population (Johnston, 1972; Wherry, 1975). Table 1 provides sample calculations demonstrating the effects of varying r_{12} and discrepant versus equal validities.

While it is possible for a beta weight to be underestimated, the gross inflation of the diagonal elements in the inverse matrix due to

Table 1

Two Predictor Regression Results with Varying Intercorrelation and Validity

Validities		Predictor	Beta Weights		Variance	Covariance	Multiple R
r_{y1}	r_{y2}	r_{ij}	$\beta_{y1.2}$	$\beta_{y2.1}$	β_i	β_{12}	
.50	.50	.00	.50	.50	0.50	0.0	.710
.50	.50	.30	.38	.38	0.68	-0.2	.620
.50	.50	.60	.31	.31	1.08	-0.65	.560
.50	.50	.90	.26	.26	3.89	-3.51	.510
.50	.50	.95	.256	.256	7.63	-7.25	.506
.50	.50	.99	.25	.25	37.64	-37.26	.501
.51	.50	.00	.51	.50	.49	0.0	.714
.51	.50	.30	.396	.381	.688	-0.2	.626
.51	.50	.60	.328	.303	1.064	-0.639	.565
.51	.50	.90	.316	.216	3.845	-3.461	.519
.51	.50	.95	.358	.159	7.563	-7.185	.512
.51	.50	.99	.75	-.246	36.11	-36.74	.511

high collinearity generally results in betas with large absolute values without regard to the true population value. While these are "correct" values their counterintuitive signs and magnitudes again have prompted the discussion of alternative estimators to OLS (Churchill, 1971; Hoerl & Kennard, 1970a; Klein & Nakamura, 1962). The notion that these inflated betas are potentially poor estimates is attributable to the effects of collinearity on the variance-covariance of the beta weights.

Equation (11) can be expressed (with $r_{12} = \alpha$) as (omitting a function of sample size)

$$(24) \qquad \mathrm{Var}\hat{\beta} = \frac{\sigma_{\varepsilon}^{2}}{(1-\alpha^{2})} \begin{bmatrix} 1.0 & -\alpha \\ -\alpha & 1.0 \end{bmatrix}$$

so that

$$(25) \qquad \mathrm{Var}\hat{\beta}_1 = \mathrm{Var}\hat{\beta} = \frac{\sigma_{\varepsilon}^{2}}{1-\alpha^{2}}$$

and

$$(26) \qquad \mathrm{Cov}\hat{\beta}_1\hat{\beta}_2 = \frac{-\alpha\ \sigma_{\varepsilon}^{2}}{1-\alpha^{2}}.$$

As multicollinearity, here expressed as α, increases it is evident that the sampling variances of the estimated coefficients increase. For example, as α increases from .5 to .9 the sampling variance increases by over 300 percent while $\alpha = .95$ gives an increment of 750 percent (Johnson, 1972). It should be noted though that poor precision in the estimation of individual coefficients does not imply that the linear combination of predictors is correspondingly poor.

This apparent anomaly is evidenced in the two predictor case with positive $\propto$ by the negative covariance of the estimates. This means that if one beta weight is overestimated, another in the same sample with which it is positively correlated with also be overestimated in absolute value but will have the opposite sign. The higher the correlation between the two variables, the more pronounced will be this tendency to compensate for errors in estimation. In the extreme case of $r_{12} = 1.0$ any pair of weights with the same sum will be exactly equivalent--for instance, weights of -4.0 and 6.0 or 3.0 and -1.0.[1] This effect is exemplified by Darlington (1968) and proven by Mason, Gunst, and Webster (1975) for the greater than two predictors case.

The above discussion has illustrated the nature of the problems cited by Johnston (1972) as being attributable to the effects of multicollinearity in the predictor set. Yet large predictor sets with, as a rule, validities above .25 or .30 are the norm rather than the exception in most MR applications. With higher validities and p greater than three or four it becomes inevitable that the deleterious effects of multicollinearity will be felt. While this problem is of minimal interest for purely predictive MR uses, it should be carefully considered when structural interpretation is the goal. In this setting the magnitudes and sampling variances of weight estimates can lead to erroneous conclusions as to the importance or predictive utility of individual variables. Thus it is important to consider ways of

[1] This is true despite the fact that perfect collinearity ($r_{12} = 1.0$) makes inversion of the predictor matrix impossible (Darlington, 1968).

detecting multicollinearity, assessing its impact, and hopefully discovering solutions to the problems it poses.

Numerous techniques have been proposed for detecting multicollinearity and the more important will be discussed.

The simplest available operational definition of unacceptable collinearity is the arbitrary establishment of a maximum permissible value for predictor intercorrelation. Aside from the arbitrariness inherent in this approach it shares the faults of the next proposal to be presented.

Klein (1962) suggests that ". . . intercorrelation or multicollinearity is not a problem unless it is high relative to the overall degree of multiple correlation . . ." (p. 101). Despite its intuitive appeal this rule of thumb is not valid. Farrar and Glauber (1967), while providing a geometric rationale for the rule, point out that perfect collinearity or the case of a completely singular predictor matrix is perfectly compatible with low pairwise correlations. A set of dummy coded contrast vectors such as commonly used for the analysis of variance whose non-zero elements exhaust the sample space would fulfill these requirements (Cohen, 1967; Cohen & Cohen, 1975). For the same reasons measures based on average intercorrelations (Cureton, 1971; Kaiser, 1968; Meyer, 1975) are inadequate warnings of severe multicollinearity.

A measure presented by Kmenta (1971) is the coefficient of determination $R^2_{(j)}$ which is obtained by regressing the criterion variable on all predictors excluding x_j. If a high degree of collinearity is present in the data the discrepancy between $R^2_{(j)}$ and the coefficient of determination for the full predictor set will be quite small. However, a small difference may simply be reflective of the worthlessness of x_j as a predictor variable. This is illustrated by Darlington's (1968) suggestion of this exact comparison for estimating the importance of individual predictors. Furthermore, this measure does not depict the nature of the collinearity, i.e., which variables are involved in the relationships.

Another measure with the same limitations as $R^2_{(j)}$ is based on the _F_ statistic obtained from fitting the full model and the _t_ statistics obtained by deleting one variable at a time from the equation. If the overall _F_ is significant and the individual _t_ tests are not, multicollinearity is indicated. However, this occurrence is unusual even with high collinearity (Mason, Gunst, & Webster, 1975), and, like the previous measure, the nature of the collinearity is not specifiable.

A single measure which summarizes the collinearity present in the entire predictor matrix, again without providing information as to its nature, is provided by the determinant, symbolized $|X'X|$. As $|X'X|$ is in standardized form, $0 \leq |X'X| \leq 1.0$ while if a linear dependence satisfying equation (19) exists the determinant is equal to zero. This measure provides at least an ordinal indicant of the presence of multicollinearity although the collinearity could be attributable to one or several very small latent roots. Under the assumption of multivariate normality (not generally tenable in the assumed fixed-X case, as discussed earlier) work by Wilks (1932) and Bartlett (1950) indicates that a chi-square test of the departure of the determinant from zero is possible. Further, the determinant obtained by deleting one variable or set of variables from the matrix forms an _F_ ratio with the determinant of the full p-order matrix. These tests are however very sensitive to departures from normality and are of sufficient complexity to discourage their frequent use (Farrar & Glauber, 1967). Much the same information can be obtained more readily by the methods to be discussed next.

Johnston's (1972, pp. 162-163) conclusion that the standard error of beta weight estimates should give adequate warning of the presence of multicollinearity can be extended to provide more exact information. The standard error of a single beta weight, $\hat{\beta}_i$, is defined to be

$$(27) \qquad \text{S.E.}\hat{\beta}_i = \sqrt{\frac{C_{ii}(1 - R^2_{y.1,2,\ldots,p})}{N - p}}$$

where C_{ii} is the diagonal element of $(X'X)^{-1}$ corresponding to the i^{th} predictor. This measure provides an intra-matrix indication of collinearity but still does not facilitate inter-matrix comparisons. A more useful measure is the C_{ii} component of the standard error formula which indicates collinearity without reference to the coefficient of determination for the full equation.

This diagonal element of the inverse intercorrelation matrix of predictors (actually a transformation of it) has been termed the variance inflation factor by Marquardt (1970) and has been employed as an indicant of multicollinearity by Marquardt and Snee (1975) and Snee (1973). This element, C_{ii}, and the off-diagonal values of the matrix can be expressed in terms of more familiar quantities to demonstrate their utility. If the symbol $s_{i.j,\ldots p}$ is used to represent the square-root of the residual variance obtained when any one predictor is regressed on the remaining p-1 predictors (i.e., $s_{i.j,\ldots p}$ =

$\sqrt{(1 - R^2_{i.j,\ldots p})}$, then C_{ii} is the reciprocal of this variance--$C_{ii} = \frac{1}{(s_{i.j,\ldots p})^2}$. This provides a convenient means of assessing multicollinearity (Farrar & Glauber, 1967) as the squared multiple correlation of each predictor regressed on those remaining is implicit in the inverse matrix of (X'X). The relationship is

$$(28) \qquad R^2_{i.j,\ldots p} = 1 - \left(\frac{1}{C_{ii}}\right) .$$

Thus if perfect collinearity exists (19), the C_{ii} element will be infinitely large and the matrix is seen as singular. For matrices which do not exactly satisfy (19), the natural range of C_{ii} is simply greater than or equal to one with equality obtaining in the orthogonal variate case. If a single high-collinearity exists (high pairwise correlations) the large C_{ii} and C_{jj} will indicate which variables are involved. However, if multiple caused collinearities are present one must look to the off-diagonal elements of $(X'X)^{-1}$ for more information. An off-diagonal element C_{ij} is defined as

$$(29) \qquad C_{ij} = \frac{- r_{ij.k,\ldots p}}{(s_{i.j,\ldots p})(s_{j.i,k,\ldots p})} .$$

The numerator is a partial correlation of an order two less than the rank of the full matrix. It may be noted that inversion of the correlation matrix provides a means of quickly obtaining all of the highest order partial correlations by the formula

$$(30) \qquad r_{ij,k,\ldots p} = \frac{- C_{ij}}{\sqrt{C_{ii}C_{jj}}} .$$

Especially in cases of non-overlapping groups of multicollinearities consideration of the diagonal and off-diagonal elements of the inverse allow one to locate the variables contributing to the problem. Marquardt (1970), Mason, Gunst, and Webster (1975), and Farrar and Glauber (1967) consider this indication of collinearity to be the best available. Gordon (1967) illustrates the effects on these values produced by varying intercorrelation and subset size.

There is yet one improvement which can be suggested to further facilitate the interpretation of multicollinear matrices. Once the presence of high collinearity has been established by means of one or several high C_{ii} values or, in a summary manner, by the existence of a near-zero determinant, one is still interested in accurately pinpointing the contributions to the problem--essentially in specifying the coefficients of equation (19). A procedure which enables this is basically the stepping stone for rank-reduction alternatives to OLS or, alternatively, as an exploratory statistical method in its own right. Eigenanalysis is basic to all expositions of principal components or factor analysis, but its utility has not been widely appreciated by users of multiple regression. Eigenanalysis is essentially a procedure for extracting from a matrix the successive vectors of weights which when applied to the original variables will produce linear combinations of maximum variance. These eigen or characteristic vectors as they are also called are subject to two conditions. First, they are restricted to unit length, i.e., $v_i \; v_j = 1.0$. Secondly, each vector must maximize the residual variance extracted from the matrix subject to the condition that it is orthogonal to all other vectors. For the case of an intercorrelation matrix (X'X), the matrix equation to be solved is

(31) $$((X'X) - \lambda I)v_i = 0$$

where λ_i represents the characteristic root or variance of its associated vector of coefficients, v_i and is obtained from

(32) $$V'(X'X)V = \lambda \text{ diagonal.}$$

Solution of this equation (Cooley & Lohnes, 1971; Finn, 1974; Tatsuoka, 1971) yields a set of p roots and a matrix of dimension p x p containing the coefficient vectors. Some attributes of these values can be of use in assessing the effects of multicollinearity. For a matrix of orthogonal standardized variates each characteristic root is equal to exactly one.

Therefore, if

$$r_{ij} = 0 \text{ for all } i \neq j$$

then, $\lambda_i = 1.0$ for all i,

and

(33) $$\sum_{i=1}^{p} \lambda_i = p, \text{ the matrix rank.}$$

If the vectors are not orthogonal the sum of the roots must still equal the variance of the full matrix--thus (33) is true in all cases. However, with correlated variates the first one or several eigenvectors extracted will exhaust much of the variance and the later eigenvalues will approach zero. In the case of perfect collinearity (19) one or more of the roots will in fact be equal to or less than zero. Thus each eigenvalue is an indicant of the degree of collinearity present in the matrix and the inflation of the sum of the reciprocals of the roots away from p provides another matrix-wide summary of the severity. Of more interest than merely another summary measure are the elements of the vectors associated with small eigenvalues. These coefficients, just as in factor analytic interpretations, show which variables are the major contributors to the definition of the vector. Thus large positive or negative coefficients in a vector with a small eigenvalue indicate which variables are contributing the most to the lack of orthogonality (Marquardt & Snee, 1975; Mason, Gunst, & Webster, 1975; Snee, 1973; Webster, Gunst, & Mason, 1971). A relationship of interest (Snee, 1973) involves an alternative method of computing the diagonal elements of the correlation matrix inverse. Because the eigenvector matrix is columnwise orthogonal and of unit length ($V'V = VV' = I$) equation (32) can be rearranged to give

$$(34) \qquad (X'X) = V\lambda_D V'.$$

Using a matrix theorem for inverses ($(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$, Dorf, 1969) it is evident that

$$(35) \qquad R^{-1} = V'\lambda_D^{-1}V,$$

and

$$(36) \qquad R_{ii}^{-1} = C_{ii} = V_{i1}^2\lambda_1{}^{-1} + V_{i2}^2\lambda_2{}^{-1} + \ldots + V_{ip}^2\lambda_p^{-1}.$$

From this equation the significance of characteristic roots less than 1.0 is immediately obvious. The basis of the standard error for any one beta weight (28) is a direct function of the spectrum of eigenvalues for the matrix upon which it is computed. An eigenanalysis and other multicollinearity statistics discussed above are presented in Table 2 based on a numerical example taken from Cooley and Lohnes (1971). With only three variables and a rather simple pattern of interdependence the source of the collinearity is readily apparent. In more complex analyses however the information provided by large loadings (-.66 and .72 in v_3) and associated small roots (.304) can be valuable.

Because of the problems with established methods of assessing multicollinearity outlined above, it is suggested that eigenanalysis be performed on any data set in which high collinearity is suspected. Inspection of the eigenvector values should allow a researcher to pinpoint likely problem variables or sets of variables. Once the severity of the multicollinearity present in a matrix has been assessed, ways should be considered for handling the deleterious effects it can have on weight estimates. Numerous methods have been presented in the statistical, sociological, and econometrics literature and several will be discussed here.

Alternatives to Ordinary Least Squares

The two basic uses to which multiple regression estimation of weighting coefficients are applied are again relevant here. The majority of alternatives to OLS (including derivations based on the OLS procedure) are directed at maximizing the sample equation's multiple R or else its expected value on cross-validation, subject to such constraints as computational ease or the availability of adequately large data samples. Thus, many alternatives are explicitly concerned only with prediction, and several in fact make structural interpretation

Table 2

Illustrative Multicollinearity and Eigenanalysis Values

$$(X'X) = \begin{bmatrix} 1.00 & .67 & -.10 \\ .67 & 1.00 & -.29 \\ -.10 & -.29 & 1.00 \end{bmatrix}$$

Determinant $(X'X) = .4987$

$$(X'X)^{-1} = C = \begin{bmatrix} 1.84 & -1.28 & -.18 \\ -1.28 & 1.98 & .44 \\ -.18 & .44 & 1.11 \end{bmatrix}$$

$$\text{Eigenvectors} = V = \begin{bmatrix} .64 & .38 & -.66 \\ .69 & .10 & .72 \\ -.34 & .91 & .20 \end{bmatrix}$$

$$\text{Eigenvalues} = \lambda_i = \begin{array}{cc} & \text{\% of Trace} \\ \begin{bmatrix} 1.768 \\ 0.927 \\ 0.304 \end{bmatrix} & \begin{matrix} 59.0 \\ 30.9 \\ 10.1 \end{matrix} \end{array} \qquad \sum_{i=1}^{p} \lambda_i^{-1} = 4.924$$

impossible either by eliminating some variables on the basis of specified criteria or by utilizing arbitrary weights. Alternatives to ordinary least squares can be briefly summarized as falling into one of the following classes: (a) some form of data augmentation; (b) rank reduction procedures; (c) utilization of weights independent of interdependence relationships present in the initial sample.

The need for data augmentation is most explicit when the available sample size is less than the number of predictor variables for which weights are to be estimated. When this situation arises, as it often does for example in medical studies involving multiple observations on a limited number of patients, there are few alternatives to simply obtaining more subjects or eliminating some variables. The latter course of action precludes structural interpretation of a full set of weights although predictive utility may not be significantly hampered. Evaluating more subjects is often prohibitively expensive if not simply impossible. If excessive collinearity attributable to an undersampling of the regions of the data domain is evidenced by either evaluation of the eigenvectors or joint variable distributions (Webster, Gunst, & Mason, 1974), little choice remains other than to acquire observations on a larger N.

Rank reduction procedures have occasionally been employed in the last mentioned case, although interpretation may be vastly complicated. These procedures may be classified basically as either based on a posteriori orthogonalization of the data vectors or on evaluation of successive partial validities as variables are included or deleted from the predictor set. Virtually all orthogonalization models attempt to eliminate specific variance (in the factor analytic sense) from the predictor intercorrelation matrix. Thus, one approach in this area is the application of a principal components analysis of the predictors (normalized eigenvector values). Based on a scree test (Cattell, 1966) or the meaningfulness of the resulting components, an arbitrary number (Jeffers, 1966; Jolife, 1972, 1973) are retained and matrix transformations are used to re-estimate variable scores for individual subjects. These scores are then used in the usual regression computations. Examples of this method have become fairly common since the advent of readily available computers to carry out the tedious matrix manipulations (Gunst, Mason, & Webster, 1975; Jeffers, 1966; Massey, 1974, Schmitt & Coyle, 1976). Variations on this approach have utilized the characteristic vectors as predictors (Gunst, Mason, & Webster, 1971), inserted communality estimates in the R matrix (Horst, 1941), and attempted to estimate the R^{-1} matrix (Guttman, 1958) rather than R itself. One method (Burket, 1964) augments the predictor matrix with the vector of criterion validities before principal axes orthogonalization. Finally, all analyses based on components or axes may also be subjected to rotation (for example, varimax or quartimax) before being used to re-estimate subjects' scores.

If a principal components analysis is employed and the number of retained components is the same as the number of original variables it can be shown that the multiple regression equation derived will be identical to that obtainable from the raw variables (Darlington, 1968; Herzberg, 1968). Therefore, only cases in which fewer factors are extracted than the number of original variables can potentially be of interest. The argument in favor of such rank reduction is usually based on the well known fact that if the variables being factored contain substantial error variance, it will tend to be concentrated in the vectors associated with small roots. Potentially serious problems exist in applications of any of these methods. The distributional theory is exceedingly complex for those analyses employing communality estimates and this leaves significance testing of derived weights a virtually intractable problem (Burket, 1964). Further it is possible that the factor accounting for the least variance in the predictor set, and therefore the prime candidate for omission, in fact correlates perfectly with the criterion (Darlington, 1968). Description of the variables re-estimated from factor matrices is also generally complicated by "intermediate" loadings, and the indeterminancy of factor scores up to a linear transformation and this in turn obfuscates efforts to interpret the subsequent regression equation.

Variable deletion based on various criteria has been proposed when degrees of freedom are limited or when collinearity is a problem (Draper & Smith, 1966). In all cases deletion procedures attempt to maximize the validity of the initial equation subject to specified constraints, and are thus not amenable to instances in which structural interpretations of a full rank predictor matrix is of interest. As Darlington (1968) notes, removing the variable with the smallest beta weight is not guaranteed to produce the equation with the highest population validity for that rank model. Accretion methods of variable selection begin with the variable having the highest zero-order validity and then in successive steps add those variables which will give the greatest increase in the multiple R for the equation (Draper & Smith, 1966). Horst and MacEwan (1960) suggest the reverse of this procedure and note that the two methods--forward selection and backward elimination--will not in general yield the same equations. Both procedures are terminated on the basis of arbitrary criteria such as validity increment or number of variables included. Stepwise regression is essentially identical to forward selection but additionally it tests at each step all variables already in the equation. If, because of the inclusion of subsequent predictors, a variable's partial correlation has fallen below a specified value it is then eliminated and the procedure continues to evaluate the remaining candidates for the equation (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975). Numerous variations on these three basic approaches have been suggested (Anderson & Fruchter, 1960; Burket, 1964; Furnival & Wilson, 1974; Hocking & Leslie, 1967; LaMotte & Hocking, 1970; Rock, Linn, Evans, & Patrick, 1970) but in general these three have been preferred.

The case of interest in the present paper is that in which one wishes, for either predictive or interpretative purposes, to obtain an equation based on all p variables upon which data were collected. Accordingly, the relative merits and problems of reduced rank procedures will not be discussed further.

The selection and application of non-least squares estimated weights has received a great deal of attention especially in the psychological decision-making literature (Einhorn & Hogarth, 1975). In general the motivation for development of alternative weighting strategies has had three facets: (a) especially when computations must be done by hand, the complexity of the work necessary to calculate OLS weights is prohibitive; (b) high collinearity situations combined with less than perfect reliability of measurement make it quite likely that arbitrary weights will outperform unstable beta estimates in subsequent usage; (c) situations in which it is desired to correlate a linear function of the predictors with a criterion but a sufficiently large sample on which to estimate beta weights is not available.

A variety of combinatorial schema have been evaluated in the literature (Claudy, 1972; Lawshe & Schucker, 1959; Trattner, 1963; Wesman & Bennett, 1959) such as raw score addition so that variables are weighted by their standard deviations, addition of standardized scores, weighting by the reciprocal of the standard deviation, and weighting by the validity coefficient. The concensus has developed that equal weights for situations in which N is less than approximately 50 are superior or only slightly inferior to OLS weights regardless of the number of predictors. The comparison made is generally between the multiple R obtained from application of the original beta weights in a cross-validational sample (a set of cases not included in the original estimation of the weights) and the multiple R produced by unit weights. A comprehensive Monte Carlo study of the empirical performance of unit weights versus sample beta weights when they are validated in the population was done by Schmidt (1971). For 40 combinations of N and p comparisons on a variety of correlation matrices sampled from the literature, he demonstrated that the maximal superiority (in terms of obtained R^2 values) of beta weights averaged over 100 samples was only .083. When suppressor variables were removed this maximum dropped to .039. Both maximum beta versus unit weight discrepancies were in fact obtained in the populations themselves where the beta weights were error free, i.e., parameter values rather than sample estimates. No other weighting scheme has been shown to be so consistently comparable to the performance of beta weights.

The results provided by Schmidt's analysis are in accord with the suggestions of Einhorn and Hogarth (1975) and Dawes and Corrigan (1974) who derived their conclusions from comparative studies of the human decision making process. Dawes and Corrigan summarize their results by stating that to obtain stable prediction equations in situations where all variables are subject to error it is necessary simply to select relevant predictors, determine their sign, make all predictors comparable, and then add. While this solution appears at first to be a panacea for the numerous problems encountered in MR Roose and Doherty (1976) noted several difficulties in attempting to apply these suggestions. They found that selecting the variables without the use of some sort of stepwise procedure an arduous task. Nor had they any manner of a priori determining the predictor signs. Those they used were based on the validity coefficients for the selected variables. In their words (Roose & Doherty, 1976), ". . . the success of unit weighting as demonstrated in the present study rested upon crutches fashioned from the very MR procedure bested by unit weighting" (p. 245). Wainer (1975, 1976) has formulated the expected loss attributable to the use of unit rather than OLS weights and noted that for practical purposes the loss is so small that the OLS procedure is not justifiable. Again though, his derivations assume some sort of selection and sign assignment a priori, no suppressors, as well as a maximal spread in the beta weights of only .5, conditions which it is frequently impossible to meet.

If one is interested in full rank multiple prediction it would appear that unit weighting is the viable alternative to MR. While structural interpretation is not possible, except on the gross level of zero weights implying no utility and weights of one or minus one indicating acceptable utility, the robustness of unit weights makes them worthy of further consideration especially in cases of extreme collinearity where large beta weight sampling variances are common.

Three methods of assessing the signs of unit weights (discussed below) are compared to OLS in this work. The problems with standard multiple regression which prompted researchers to consider unit weighting and various orthogonalization schema have recently given rise to a modified OLS methodology. The details of this method, termed ridge regression, are discussed next.

Ridge Regression

Hoerl (1962) originally proposed this modified regression method specifically to deal with the problems of severe multicollinearity discussed above. In exemplifying the method of ridge analysis (Hoerl & Kennard, 1970a; 1970b) the errors associated with non-experimentally collected data are noted in that (X'X) is not nearly a unit or identity matrix. Weighting coefficients derived from such a matrix are often of incorrect sign and have inflated values, as was noted before. The undesirable nature of such weights are expressed by Hoerl and Kennard (1970a);

> . . . the least squares estimates [which] often do not make sense when put into the context of the physics, chemistry, and engineering of the process which is generating the data. In such cases, one is forced to treat the estimated predicting function as a black box or to drop factors to destroy the correlation bonds among the x_i used to form X'X. Both these alternatives are unsatisfactory if the original intent was to use the estimated predictor for control and optimization (p. 55).

The suggestion offered in such cases is the use of

(37) $$\hat{\beta} = [X'X + kI_p]^{-1}X'y, \ k \geq 0$$

for estimating beta weights rather than equation (8). This procedure it is claimed modifies the weight values such that they are less extreme in absolute value and thus necessarily have reduced variance. The technique can also be used to generate a trace of the effects of increasing k values on the coefficients which portrays the differential effects on each. Hoerl and Kennard (1970a) contend that by reducing the variability of the coefficients a more accurate estimate of the parameter values can be obtained although the resultant estimates are biased.

The derivation of this approach considers the variance of the coefficient vector $\hat{\beta}$ (equation 11) and notes that the expected value of the squared distance (L^2) from $\hat{\beta}$ to β is

(38) $$E(L^2) = \sigma^2 \text{Trace } (X'X)^{-1}$$

or equivalently,

(39) $$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 \text{Trace}(X'X)^{-1}.$$

Hoerl and Kennard also demonstrate that (38) is equivalent to

(40) $$E(L^2) = \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1}.$$

The lower bound for the average squared distance between the sample coefficients and the parameters is given by σ^2/λ_{min}. This corresponds to the previous discussion of multicollinearity wherein it was noted that small eigenvalues are one of the best indicants of unstable weights. The authors suggestion of augmenting the diagonal of (X'X) with small positive quantities ($0 \leq k \leq 1$) has the effect of decreasing the diagonal elements of the predictor inverse matrix. This in turn deflates the absolute values of the beta weights and reduces their collective variance.

Hoerl and Kennard (1970a, p. 60) demonstrate that the expected squared distance from $\hat{\beta}$ to β is composed of two elements--the total variance of the parameter estimates and the square of the bias introduced by the non-least squares computation (equation 37). Thus when $k = 0$ and OLS are calculated the bias is zero. The authors show with an existence theorem that it is possible to select k greater than 0, take a little bias, and without greatly inflating the residual error variance for the equation, obtain β estimates with substantially lower mean square errors (L^2).

The problem is then one of selecting an optimal k value for use in any one matrix. The suggestion of Hoerl and Kennard (1970b) is to use a graphic display of the effects of increasing k and note the point at which four conditions are met: (a) the characteristics of the graph will be those of an orthogonal system, (b) coefficients will have reasonable absolute values, (c) coefficients with apparently incorrect signs at $k = 0$ will have changed to correct signs, and (d) the residual sums of square will not have inflated to an unreasonable value. The plot (Hoerl & Kennard, 1970b, Figure 2, p. 72) shows the values of each of 10 coefficients plotted against the value of k in equation (37) which produced them. They would advocate selecting the beta weights produced by the equation with a k of approximately .25--that is, after the point of maximum decline in absolute value is passed and the coefficients are seen to be visually stable. Assessments which Hoerl and Kennard (1970b) make on the basis of this graph exemplify the utility of the procedure:

> (i) The coefficients from the ordinary least squares are undoubtedly overestimated. At least, they are collectively not stable. It is unlikely that another set of y's would give β_i like these. Moving a short distance from the least squares point $k = 0$ shows a rapid decrease in absolute value of at least two of them, namely, those for factors 5 and 6. Figure 2 shows the decrease in the squared length of the coefficient vector with k. When $k = .1$, it is 43.3% of its original value; for an orthogonal system it would be 83%.
>
> (ii) Factor 5 has the negative coefficient with the largest value. But the addition of $k > 0$ quickly drives it toward zero and it then becomes positive. Such action should not be surprising, especially when it is compared with the action of factor 6. Factor 6 also decreases rapidly but stabilizes and does not go down to zero. Factors 5 and 6 have a simple correlation coefficient of 0.84 which says that to a first approximation, they are the same factor but with different names. It would be surprising if their true effects were opposite in sign. (Without a knowledge of the underlying technology, no definitive statement can be made.) The covariance of -4.33 is driving them apart so that they are opposite in sign. The phenomenon observed here is not atypical. Positive coefficients for highly correlated factors can be stable as a sum, especially when they are correlated to various degrees with other factors.
>
> (iii) The correlations with other factors causes factor 1 to be underestimated. At $k = 0$ factor 1 is the second least important negative factor. But with the addition of $k > 0$ it increases in absolute value. The other negative factors are slightly overestimated and when sufficient $k > 0$ has been added to stabilize the system, factor 1 becomes the most important negative factor.
>
> (iv) Factor 7 is overestimated and is driven toward zero.
>
> (v) At a value of k in the interval (0.2,.3) the system has stabilized and coefficients chosen from a k in this range will undoubtedly be closer to β and more stable for prediction than the least squares coefficients or some subset of them (pp. 71-72).

Several authors have utilized the ridge regression (RR) technique. Churchill (1975) used 3001 cases selected in samples of size 50 and calculated ridge coefficients ($\hat{\beta}$) for 13 predictors. His results demonstrated a departure from the parameter values 1.7 times higher for OLS as opposed to RR. Vinod (1976) who used a modified RR method which selected arbitrary k values based on rank reduction analyses, Marquardt and Snee (1975), McDonald and Schwing (1973), and Snee (1973) all reported superiority of RR over OLS.

Several researchers have attempted to develop point estimates of k (Baldwin, 1975; Hoerl & Kennard, 1976; Lawless & Wang, 1976; McDonald & Galarneau, 1975; Newhouse & Oman, 1971) but these attempts uniformly assume that k is non-stochastic (Coniffe & Stone, 1973; Smith, 1976)--an unadmissable assumption. Further, these more exhaustive studies do not invariably demonstrate RR as superior to OLS. Thus while virtually all investigators consider RR to be an instructive mode of analysis most contend that it is preferable to OLS in all nonorthogonal situations, it still remains as much an art (selecting k) as a science. The most practical suggestion is probably that of Marquardt's (1970) variance inflation factor (VIF) which was mentioned earlier. This value for the i^{th} predictor is the i^{th} diagonal element of the matrix $[(X'X)_k^{-1}(X'X)(X'X)_k^{-1}]$. Just as the diagonal elements of $(X'X)^{-1}$ in standard form range from one to infinity as collinearity increases, so do these VIF values. Marquardt's suggestion is that k be selected at the point where these values are ". . . reasonable, certainly less than 10 . . ." (p. 609). Evaluation of the VIF along with the eigenvector weights associated with small eigenvalues (Snee, 1973; Webster, Gunst, & Mason, 1974) would appear to be the most reasonable way of ascertaining which VIF's should be deflated the most and therefore which k value should be selected.

The research reported here proposed to evaluate the relative efficiency of ridge regression as compared with ordinary least squares.

In addition, unit weighting was contrasted with these methods both because of its demonstrated utility and because it should in fact be most efficient in exactly the high collinearity situations for which RR was proposed (Wainer, 1976; Wainer & Thissen, 1976).

CHAPTER III

METHOD

Consideration of the possible approaches to these comparisons favored a Monte Carlo study in which the sample size and collinearity could be controlled. Accordingly three matrices were selected from the literature. The factor structures for each of these matrices were input to the Ohio State Correlated Score Generation Program (Wherry, Naylor, Wherry, & Fallis, 1965) which produces multiple random samples corresponding to the structure. Generated samples from each of the three matrices were pooled to form three populations of 6000 cases each. Each matrix selected had 10 predictors so that a total of 165 coefficients were estimated. The maximum obtained discrepancy between a target and an estimated population correlation was .031.

The Population Matrices

The first matrix selected was used as an example by Hoerl and Kennard (1970a) and was taken from Gorman and Toman (1966). This matrix (hereafter referenced as HOPOP, Table 3) was selected both because of its previous use as an RR supportive example and because of its broad range of predictor intercorrelations. Two other matrices were selected so as to broaden the scope of the comparisons. These matrices were considered more typical of those generally encountered in psychological

Table 3

Population Matrix Based on 6000 Cases--HOPOP[a]

Variable	1	2	3	4	5	6	7	8	9	10
2	-.034									
3	.511	.003								
4	.116	-.166	-.018							
5	-.713	.052	-.610	-.065						
6	-.870	.086	-.659	-.085	.846					
7	-.083	.242	-.031	.006	.371	.123				
8	-.003	.006	.337	.078	-.370	-.198	-.502			
9	-.111	.113	-.074	.003	-.123	.050	.075	-.172		
10	-.360	-.299	-.441	-.093	.542	.442	.412	-.462	.031	
Criterion	-.816	-.104	-.640	-.102	.579	.814	.040	.058	.156	.450
Population Beta Weights	-.175	-.226	-.371	-.109	-.459	.811	.289	.384	.080	.092
Multiple R = .947		R^2 = .897	Determinant = .0034							
Eigenvalues	3.709	1.554	1.313	1.041	.952	.657	.357	.215	.132	.069

[a] Taken from Hoerl and Kennard (1970b).

and sociological applications. Table 4 illustrates the high average intercorrelation matrix reproduced from the factor structure of a matrix employed by Rock, Linn, Evans, and Patrick (1970) and originally taken from Klein and Evans (1969). Table 5 is the low average intercorrelation matrix used by Rock et al. (1970) and taken from Klein and Evans (1968). These two matrices (HIPOP and LOPOP respectively), incorporating two other predictors which were deleted for the present research, were selected by Rock et al. (1970) to evaluate four methods of predictor selection because of their representativeness. It was felt that these three data sets constituted a reasonable sample from the domain of possible matrices of interest to researchers in the social sciences. The HOPOP matrix with its negative intercorrelations and validities is atypical of most psychological data but does characterize occurrences in the economics and management literature. Additionally its use by Hoerl and Kennard (1970a) as an RR example without benefit of comparison with other techniques warrants its inclusion. Eigenanalyses of the HIPOP and LOPOP matrices (Table 6) illustrate their salient features. Both data sets differ from HOPOP in that their ranges of intercorrelation are more restricted typifying the data encountered in psychological and measurement studies. The first eigenvalue of the HIPOP matrix accounts for 67 percent of the total variance while the first four roots of the LOPOP data set account for only 63 percent of its variance. Thus, by any accepted definition, the HIPOP matrix would be considered highly multicollinear while the LOPOP matrix is less severely afflicted. The fact that six of its roots combined account for less than 37 percent of the possible variance however indicates that weight estimation is likely to be adversely affected.

Table 4

Population Matrix Based on 6000 Cases--HIPOP[a]

Variable	1	2	3	4	5	6	7	8	9	10
2	.582									
3	.664	.702								
4	.668	.660	.790							
5	.673	.645	.764	.749						
6	.640	.654	.731	.786	.708					
7	.659	.694	.800	.794	.766	.773				
8	.737	.574	.651	.631	.610	.653	.652			
9	.485	.461	5.29	.542	.510	.531	.624	.454		
10	.458	.538	.615	.568	.552	.552	.678	.439	.552	
Criterion	.549	.577	.627	.600	.566	.620	.611	.611	.401	.417
Population Beta Weights	.000	.145	.169	.045	.012	.166	.075	.249	-0.10	-.044
Multiple R = .709		R^2 = .502	Determinant = .0003							
Eigenvalues	6.747	.744	.552	.418	.382	.323	.249	.232	.181	.174

[a] Taken from Rock, Linn, Evans, and Patrick (1970).

Table 5

Population Matrix Based on 6000 Cases--LOPOP[a]

Variable	1	2	3	4	5	6	7	8	9	10
2	.266									
3	.304	-.049								
4	.235	.090	.553							
5	.156	.095	.463	.495						
6	.175	.115	.386	.416	.357					
7	.119	-.026	.338	.222	.189	.189				
8	.049	-.033	.111	.126	-.023	.128	.098			
9	.030	.017	.099	.114	.141	.079	.038	.035		
10	.189	-.008	.532	.430	.281	.262	.289	.208	.071	
Criterion	.138	-.057	.352	.171	.238	.218	.193	.086	.113	.309
Population Beta Weights	.050	-.066	.208	-.137	.107	.093	.055	.023	.070	.168
Multiple R = .428		R^2 = .179	Determinant = .1625							
Eigenvalues	3.051	1.236	1.047	.078	.860	.767	.658	.583	.453	.367

[a] Taken from Rock, Linn, Evans, and Patrick (1970).

Table 6

Eigenvectors of the Population Matrices

Variable					HIPOP					
1	.31	.38	-.38	-.23	.09	.34	-.57	-.29	.05	-.17
2	.31	.05	.37	-.16	-.84	.07	-.12	.04	-.07	.09
3	.34	.03	.23	.05	.09	.12	.39	-.36	.71	-.09
4	.34	.06	.16	.34	.19	-.18	-.06	-.46	-.37	.56
5	.33	.09	.17	.24	.25	.54	.01	.63	-.02	.19
6	.33	.09	.10	.30	.07	-.65	-.37	.30	.21	-.30
7	.35	-.13	.08	.03	.08	.09	.32	-.13	-.53	-.66
8	.30	.42	-.40	-.37	-.03	-.31	.48	.23	-.08	.21
9	.26	-.55	-.64	.34	-.28	.06	.04	.02	.11	.09
10	.28	-.58	.16	-.63	.29	-.11	-.18	.07	.02	.16
					HOPOP					
1	-.41	.36	-.14	.07	.02	-.36	.15	.03	.60	-.40
2	.01	.10	.77	.09	.33	-.22	.29	.37	-.04	.11
3	-.39	.14	.07	.20	.02	.61	-.42	.48	.08	-.02
4	-.06	-.02	-.36	-.50	.76	.02	.01	.19	-.05	.04
5	.47	-.06	.00	.16	.18	.03	-.26	-.03	.69	.41
6	.46	-.28	.11	-.04	.05	.05	-.19	.20	.05	-.78
7	.21	.59	.12	.14	.30	.40	.09	-.52	-.09	-.19
8	-.25	-.57	.09	.07	.12	.43	.51	-.24	.28	-.07
9	.04	.17	.30	-.80	-.36	.22	.02	-.03	.25	.04
10	.36	.24	-.36	.09	-.21	.22	.59	.48	.01	.06
					LOPOP					
1	.24	.46	.35	.01	-.36	.43	.46	.15	.07	.23
2	.07	.73	.22	-.14	.01	-.34	-.46	-.11	-.01	-.23
3	.46	-.13	-.04	.14	-.08	.24	.04	-.03	.02	-.82
4	.44	.03	-.11	.03	.22	.10	-.15	.12	-.80	.23
5	.38	.12	-.40	.11	.20	-.06	-.16	.57	.49	.18
6	.36	.09	-.04	-.04	.43	-.39	.55	-.45	.13	.07
7	.27	-.25	.14	.19	-.61	-.64	.04	.14	-.08	.09
8	.13	-.32	.61	-.54	.29	-.07	-.02	.35	.07	-.06
9	.11	.02	-.48	-.79	-.35	.02	.05	-.10	-.02	-.01
10	.39	-.23	.17	.02	-.10	.26	-.47	-.52	.29	.33

[a] Each population contained 6000 cases.

Samples

Twenty-five samples of sizes 30, 60, 90, 120, and 200 were drawn from each population using a random sampling procedure available in the SPSS package (Nie et al., 1975). These sample sizes were selected to span the range of values for which unit weights have been demonstrated to be superior to OLS (Schmidt, 1971). Each sample (375 in all) was standardized and input to a program written by the author for the necessary least squares, unit weights and ridge regression computations.

Equation Estimation

Five equations were estimated in each of the samples. OLS weights and the multiple R they produced were calculated according to equation (8). Three unit weight equations were also estimated in each sample. The first equation was produced by assigning the sign of the fallible sample beta weights (BU) to the unit weighting coefficients. The sign of each validity coefficient (VU) in each sample was also used to determine the unit weight signs. Third, the infallible population beta weight signs (PU) were employed. This third method implies that the investigator has prior information as to the correct signs, presumably on the basis of previous experience with the variable. These three methods were selected because, in an applied situation, they correspond to the manner in which one would generally determine the unit signs.

For the ridge regression equation in each sample the value of k in equation (37) was determined by the following rule: select the largest k possible (stepsizes of .01) with the restriction that no diagonal element of $[X'X)_k^{-1}(X'X)(X'X)_k^{-1}]$ is less than 1.0. Attendant to the earlier discussion of VIF's (Marquardt, 1970; Snee, 1973) pilot work was done experimenting with a variety of selection rules based on these values. It was found that the above rule always selected reasonable k values at a point just slightly lower than a visual trace examination would suggest. This criterion is also in keeping with more recent analytical attempts to define k which have generally found that less bias (i.e., small k values) can adequately handle the problems of multicollinearity (Guilkey & Murphy, 1975; McDonald & Galarneau, 1975).

Data Analysis

Virtually all studies evaluating ridge regression to date have, at least implicitly, been concerned only with structural interpretation. Previous Monte Carlo studies (Hoerl, Kennard, & Baldwin, 1976; McDonald & Golarneau, 1975) which had available the true parameter values of β based their evaluations on the mean square error (MSE) criterion, i.e., $\sum_{i=1}^{p} (\hat{\beta}_i - \beta_i)^2$ with the $\hat{\beta}_i$ vector being produced by either OLS or RR. While this comparison statistic accurately reflects the average precision of the β_i point estimates, it does not provide for assessment of the predictive utility of the overall linear combination. It is possible that while one method of estimating β_i will have a lower MSE than another, the predictive utility of the latter will be superior.

Because of this consideration the predictive ability of all five equations was evaluated as well as the MSE. As the RR procedure necessarily decrements the coefficient of determination as compared to that of OLS in the estimation sample, these initial R and R^2 values were evaluated. Due to the overfitting of the regression surface in the estimation sample discussed earlier, a practical measure of an equation's utility is its performance in a cross-validation sample. However, as Schmidt (1971) has noted, a researcher is not interested in how a set of weights do in a single random replication sample but rather in how they perform in the long run, i.e., how they compare with the predictive utility of the infallible population weights. Accordingly, the equations estimated for each sample were cross-validated in the populations from which they were drawn. The formula for the cross-validated multiple R is (Nunnally, 1967)

$$(41) \qquad R_w = \frac{w'X'y_{pop}}{\sqrt{(w'(X'X)_{pop}w)}}$$

where (w) is the appropriate vector of unit, RR, or OLS weights.

The final comparison statistic, like MSE, is applicable only to the RR and OLS results. The coefficient of variation proposed by Churchill (1975) is calculated by dividing the square root of the MSE for a single coefficient by the true parameter value. These coefficients of variation (CV) can then be averaged over predictors, sample sizes, and/or populations for summary purposes.

CHAPTER IV

RESULTS AND DISCUSSION

Estimation sample results for the five equations discussed above are presented in terms of the obtained mean coefficients of determination in Table 7. As noted above, the bias factor due to the k value in equation (37) results in a lower R^2 for ridge regression than ordinary least squares in all cases. The average values of these differences for 25 samples are presented for all sample sizes and for the three populations in Table 8. The magnitude of the positive values in Table 8 reflect the higher average obtained R^2 values for OLS over 25 samples for each of the four alternative weighting methods. It should be noted that the lower obtained values for RR equations (approximately .038) will in general make them better estimates of the population cross-validity due to the overfitting of sampling error present in the estimation sample. Possible exceptions to this conclusion are cases in which either an OLS or RR initial equation cross-validates in a single instance upward in terms of the R^2. This in fact occurred, on the average, for samples of sizes 90, 120, and 200 drawn from the HOPOP matrix and estimated by RR. The special nature of this population will be discussed below. A final difference to be noted between OLS and RR in Table 7 is the reduced range of R^2 estimates provided by RR over

Table 7

Initial R^2 - LS, RIDGE, BU, VU, PU[a]

		Sample Size				
Population	Equation	30	60	90	120	200
	LS	.663	.615	.578	.544	.527
	RIDGE	.587	.562	.540	.513	.502
HIPOP	BU	.415	.478	.468	.463	.463
	VU	.466	.485	.486	.471	.470
	PU	.483	.483	.493	.477	.476
	LS	.933	.918	.911	.907	.902
	RIDGE	.869	.864	.857	.858	.856
HOPOP	BU	.641	.672	.683	.692	.696
	VU	.688	.642	.636	.651	.674
	PU	.698	.700	.700	.697	.696
	LS	.434	.307	.265	.228	.208
	RIDGE	.395	.291	.257	.222	.204
LOPOP	BU	.241	.198	.191	.161	.158
	VU	.238	.184	.169	.154	.146
	PU	.125	.149	.150	.150	.155

Note. All entries are mean values based on 25 samples.

[a] LS = ordinary least squares; RIDGE = ridge regression; BU = unit weights with signs determined by sample beta weights; VU = unit weights with signs determined by sample validity coefficients; PU = unit weights with signs determined by infallible population beta weights.

Table 8

Mean Initial R^2 Superiority of Least Squares

Over RIDGE, BU, VU, PU[a]

		Sample Size					
Population	Equation	30	60	90	120	200	Average
HIPOP	RIDGE	.076	.053	.038	.031	.025	.045
	BU	.248	.137	.1o0	.081	.063	.128
	VU	.197	.130	.092	.071	.057	.109
	PU	.180	.132	.085	.067	.051	.103
HOPOP	RIDGE	.064	.054	.054	.049	.046	.053
	BU	.292	.246	.228	.215	.206	.237
	VU	.245	.276	.275	.256	.228	.256
	PU	.235	.218	.211	.210	.206	.216
LOPOP	RIDGE	.039	.016	.008	.006	.004	.015
	BU	.193	.109	.074	.067	.050	.099
	VU	.196	.123	.096	.074	.062	.110
	PU	.305	.158	.115	.078	.053	.142
AVERAGE	RIDGE	.060	.041	.033	.029	.025	.038
	BU	.244	.164	.137	.121	.106	.155
	VU	.213	.176	.154	.134	.116	.159
	PU	.240	.169	.137	.118	.103	.154

Note. All entries are mean values based on 25 samples.

[a] RIDGE = ridge regression; BU = unit weights with signs determined by sample beta weights; VU = unit weights with signs determined by sample validity coefficients; PU = unit weights with signs determined by infallible population beta weights.

different sample sizes. It appears then that RR is somewhat less sensitive to the size of the sample in which weights are estimated than is OLS. Over the three populations and five sample sizes, the range of ridge estimated coefficients of determination are approximately 37 percent less than those of OLS estimates. All three unit weight equations demonstrate the same relative indifference to sample size and in some cases to be discussed below, provide better estimates of actual utility than either OLS or RR.

For predictive purposes the R^2 obtained in the initial sample is typically not of interest beyond indicating whether the linear combination of predictor variables has any utility at all. Cross-validated (typically in only a single holdout sample) or formula estimated coefficients of determination are the usual criteria for utility decisions. In general, the latter approach has been shown to be preferable (Schmitt, Coyle, & Rauschenberger, 1977); however in a Monte Carlo study such as considered here, one has available the actual population matrix which obviates the need for estimates of long-term cross-validated efficiency. Table 9 presents the results, for the four relevant equation types, of applying the sample estimates to the population from which the data were drawn. As this step concerns validation of sample dependent values the unit weight equations signed by the infallible population beta weights (PU in Table 7 and 8) are not evaluated. Table 10 contains the average differences between OLS and the RR equations, the unit weights signed by the sample validity coefficients (VU) and the weights signed by the sample beta weights (BU). Negative entries in this Table (10) indicate that the equation in question obtained a higher average cross-validated R^2 than did OLS for the same population and sample size.

Table 9

Cross-Validated R^2 - LS, RIDGE, BU, VU[a]

		Sample Size				
Population	Equation	30	60	90	120	200
HIPOP	LS	.345	.405	.447	.466	.481
	RIDGE	.428	.456	.474	.482	.489
	BU	.231	.326	.358	.390	.412
	VU	.465	.465	.465	.465	.465
HOPOP	LS	.845	.872	.879	.886	.890
	RIDGE	.838	.858	.867	.875	.880
	BU	.574	.640	.673	.684	.697
	VU	.601	.566	.588	.612	.659
LOPOP	LS	.050	.085	.111	.130	.147
	RIDGE	.056	.089	.113	.132	.148
	BU	.040	.056	.075	.092	.113
	VU	.093	.116	.124	.128	.135

Note: All entries are mean values based on 25 samples.

[a] LS = ordinary least squares; RIDGE = ridge regression; BU = unit weights with signs determined by sample beta weights; VU = unit weights with signs determined by sample validity coefficients.

Table 10

Mean Cross-Validated R^2 Superiority of Least Squares

Over RIDGE, BU, VU[a]

		Sample Size					
Population	Equation	30	60	90	120	200	Average
	RIDGE	-.083	-.051	-.027	-.016	-.008	-.037
HIPOP	BU	.144	.079	.089	.076	.069	.085
	VU	-.120	-.060	-.018	.001	.016	-.036
	RIDGE	.007	.014	.012	.011	.010	.011
HOPOP	BU	.271	.232	.206	.202	.193	.221
	VU	.244	.306	.291	.274	.231	.269
	RIDGE	-.006	-.004	-.002	-.002	-.001	-.003
LOPOP	BU	.010	.029	.036	.038	.034	.029
	VU	-.043	-.031	-.013	.002	.012	-.015
	RIDGE	-.027	-.014	-.006	-.002	.000	-.010
Average	BU	.132	.113	.110	.105	.099	.112
	VU	.027	.072	.087	.092	.086	.073

Note. All entries are mean values based on 25 samples.

[a] RIDGE = ridge regression; BU = unit weights with sign determined by sample beta weights; VU = unit weights with signs determined by sample validity coefficients.

Inspection of Tables 9 and 10 shows that BU equations are generally the poorest while RR provides the best average results. These results are more evident if one ignores the obtained values for the Hoerl and Kennard (1970b) population. In the high and low intercorrelation populations, ridge regression outperforms OLS by a small margin (.02) for all sample sizes. In the HOPOP matrix the situation is reversed with OLS demonstrating a slight superiority (.01) over RR.

As Tables 7 through 10 concern predictive utility rather than structural interpretation, it is at this point that the efficiency of the various unit weighting schemes must be considered. Schmidt (1971) noted that with simulated data such as presented here, violations of the assumptions of multiple regression (linearity, homogeneity, and normality of conditional variances) cannot occur. Such violations apparently occur in approximately 20 percent of actual empirical data sets (Sevier, 1957; Schmidt, 1971; Tupes, 1964) and their effect is to attenuate the predictive utility of OLS. Therefore, in this simulation differences between OLS obtained R^2 values and those of unit weights should be taken as maximal estimates. In practice, OLS will be somewhat less efficient than is indicated here.

In the HIPOP and LOPOP matrices (Table 9 and 10) the results for unit weighting methods are similar to those reported by Schmidt (1971). As concluded in that study, when no suppressor effects are present, a sample size of approximately 180 is necessary before OLS will demonstrate a distinct superiority over unit weights. In Table 9 the high and low intercorrelation populations show OLS to be useful upon cross-validation in the range between 120 and 200 cases. It is also concluded from these tables (9 and 10) that signing unit weights with the sign of the sample beta weight estimate is not generally advantageous. This is congruent with Hoerl and Kennard's (1970a, 1970b) rationale for RR; that is, that when collinearity is high, the sample beta weights will frequently exhibit incorrect signs and indicate excessive suppressor effects. Thus, when previous experience with the variables permits one to decide the sign of the unit weight for each predictor, these signs should be employed. This method is conceptually at least, preferable to both the BU and VU sign assignment as it is independent of sampling fluctuations.

In practice many uses of MR involve variables (as predictors or criteria) for which one could not confidently decide on their sign before analysis (Roose & Doherty, 1976). The conclusion to be drawn from this study is that the next best alternative is to use the sign of each predictor's zero-order validity coefficient.

The Hoerl and Kennard (1970b) population matrix (HOPOP in Tables 7 through 10) presents several contradictions to the above mentioned conclusions. This population is not typical of those encountered in social science data; generally, its coefficient of determination is higher than the norm, four of the validities are negative, and five variables in the population are identified as suppressors (see Table 3). It is ironic that this matrix was chosen by Hoerl and Kennard (1970b) as an example of the advantages of RR over OLS. Across all sample sizes investigated in this study RR equations based on random samples from the HOPOP matrix are dominated by OLS. Ordinary least squares also demonstrates higher cross-validated coefficients of determination than do either beta weight or validity signed unit weights. With regard to the RR results it must be concluded that the biasing factor of equation (37) "overcorrected" the weights of some predictors in this population and thus reduced the cross-validated R^2. This occurrence emphasizes the need for an analytical determination of an optimal biasing parameter (k) which ideally could adopt different values for different predictors. This would seem to be indicated as advantageous for sample data from a matrix such as HOPOP where the collinearity is not uniform across the predictors. A mixture of high and low pairwise intercorrelations (Table 3) presumably requires a variable bias factor. This conclusion is supported by the results for the HIPOP and LOPOP matrices both of which demonstrated RR as superior to OLS upon cross-validation of the sample equations in the population. These discrepancies further demonstrate that the determinant is an insufficient indicant of the degree of collinearity insofar as its value might be used to determine whether OLS or RR should be applied. The LOPOP population actually has a determinant 48 times larger (indicating less severe multicollinearity) than the HOPOP matrix, yet RR was superior on the LOPOP samples and not on the HOPOP samples.

Conclusions as to the predictive utility of these various combinatorial schema would seem to be as follows:

1. In agreement with Schmidt (1971), unit weights should be employed in samples of under approximately 200 cases.
2. In the absence of prior knowledge, validity coefficients should be used to determine the sign of each predictor's unit weight.
3. The predictive utility of ridge regression, while superior to OLS and unit weights in some instances, would not seem to be great enough to warrant its use. If an analytic determination of the bias parameter k is developed, ridge regression would

seem to be practical for analyses in which the intercorrelation is both "high" and consistent throughout the matrix and sample size is not very large. While this study is not conclusive, it appears that RR is most useful for predictive purposes in the same sample size range as are unit weights.

The focus of the discussion now turns to consideration of the accuracy of weight estimation by the OLS and RR methods as opposed to the predictive utility of their respective linear combinations. Table 11 presents the average mean square error (MSE) of estimation values, the average bias factor (k), and the calculated average sum of reciprocals of the eigenvalues for each population and sample size. It should be recalled that the selection of the value for k determines, along with sample specific collinearity, the value which will result for MSE. Thus, as long as an analytic solution for the bias factor is not available, individuals may rightly argue for the appropriateness of values other than those employed here. It is considered, however, that the method of determining k employed in this study yields reasonable results which are consistent with published uses of RR. Further, as noted by Churchill (1975), a Monte Carlo study is potentially susceptible to the criticism of optimizing the selection of the bias factor so as to conform to the population specifications. Thus, it is argued that the arbitrariness of a "reasonable" selection rule such as used herein will permit greater generalizability of results as we await a solution to the problem of analytically optimizing k on the basis of sample information only.

The mean square error values in Table 11 can be interpreted as a summary measure of the accuracy with which weights were estimated.

Table 11

Equation Mean Square Errors

		Population								
Sample Size	Equation[a]	HIPOP			HOPOP			LOPOP		
		MSE	$\Sigma\lambda^{-1}$	k	MSE	$\Sigma\lambda^{-1}$	k	MSE	$\Sigma\lambda^{-1}$	k
30	LS	.085	51.15		.020	54.46		.064	20.47	
	RIDGE	.023	23.67	.15	.041	27.04	.08	.043	16.56	.07
60	LS	.489	125.34		.009	43.78		.027	16.45	
	RIDGE	.033	23.01	.15	.029	26.85	.06	.022	14.75	.05
90	LS	.023	36.15		.006	39.48		.016	15.54	
	RIDGE	.009	20.87	.14	.025	25.64	.06	.014	14.52	.03
120	LS	.013	34.75		.004	38.89		.010	15.19	
	RIDGE	.006	20.43	.14	.021	26.42	.05	.009	14.27	.03
200	LS	.008	32.80		.002	37.05		.006	14.43	
	RIDGE	.004	19.87	.15	.018	26.23	.05	.005	13.84	.02

Note. All entries are mean values based on 25 samples.

[a] LS = ordinary least squares; RIDGE = ridge regression; $\Sigma\lambda^{-1}$ = sum of the reciprocals of the sample eigenvalues; k = average value of k in the expression $((X'X) + kI)^{-1}X'y$.

MSE is equal to the sum of the deviations squared about the parameter beta weight plus the squared bias. Thus, the smaller the MSE value is for a particular cell, the better the average beta estimate was when one averages errors over the 10 coefficients in each sample and the 25 samples per cell. It is seen in Table 11 that RR had a smaller MSE than OLS in all HIPOP and LOPOP samples. This improvement in accuracy seems to be inversely related to the sample size available for estimation, similar to the results for predictive utility presented above. While it was concluded earlier that unit weights were preferable to RR estimates for predictive use when sample size is less than approximately 180, the same is not true here. Structural interpretation of regression estimates makes explicit the intent to characterize a system or process as a function of the magnitude of weight estimates. The substitution of arbitrary weights (i.e., unit weights) may not deter predictive use of the system's indicators but it necessarily eliminates the possibility of assessing their individual utilities.

The improvement in MSE attributable to RR is substantial in most cases. The outlying value of .489 for the least squares mean value of MSE at a sample size of 60 is attributable to one random sample's extreme beta estimates. Omitting this one sample and calculating the same statistics for OLS on 24 samples yields MSE = .033; $\Sigma\ \lambda_i^{-1}$ = 39.36 and k for RR remains unchanged at .15. It is interesting to note that this one sample's extreme beta estimates were adequately handled by the RR technique using the decision rule for k selection discussed above.

Over all sample sizes in the high intercorrelation maxtrix, the MSE due to use of RR weight estimates is approximately 91 percent less than that generated by OLS (the value is 65 percent if the one aberrant sample from the sample size 60 cell is removed). In the LOPOP matrices RR is 24 percent more accurate overall. In the HOPOP matrices, as was the case with predictive utility, RR is dominated at all sample sizes by OLS. For these samples OLS is 69 percent more efficient than RR. The conclusion is therefore similar to that for predictive considerations: for the appropriate matrices (high collinearity which is consistent across the matrix) ridge regression can provide improved weight estimation, especially for small sample sizes. Even in subjectively low intercorrelation cases (LOPOP) RR will not be worse than OLS although the extra computational labor may not be worth the slight gain in estimation precision.

Table 11 also lists the values computed for the sum of the reciprocals of the eigenvalues with and without the biasing factor (OLS and RR solutions respectively). This value was assessed by Hoerl and Kennard (1970b) as an indication of the degree to which orthogonalization had been achieved by the RR technique. If the predictors utilized had in fact been uncorrelated, this value, as demonstrated earlier, would equal 10.0 or equivalently, p, the number of predictors. RR over all HIPOP matrices demonstrated a 61 percent reduction (44 percent with the above noted sample omitted from the sample size 60 cell) in the sum of eigenvalue reciprocals as compared to OLS. In the LOPOP matrices the figure was 10 percent while in the HOPOP samples the reciprocals were 38 percent smaller. These results demonstrate the inappropriateness of this value (the sum of reciprocals of eigenvalues) as an indicant of the utility of the RR technique. As noted for the HOPOP matrices reduction in the size of this value is an artifact of the application of equation (37) and does not necessarily imply either enhanced precision of estimation or superior predictive ability.

Churchill's (1975) modified coefficient of variation can also be used to assess the advantages of one technique relative to another. This value (CV) is calculated by dividing the square root of an estimate's MSE by the absolute value of the parameter it is intended to estimate. These values can then be averaged for summary purposes. Table 12 presents the ratio of the average CV produced by OLS to that of RR. Again, deleting the single outlying sample from the HIPOP 60 cell reduces the value reported in Table 12 to approximately 1.71. These results are consistent with the conclusions drawn on the basis of MSE comparisons; RR is substantially more accurate at small sample sizes with high, consistent collinearity than is the OLS technique. This dominance diminishes as sample size increases and it is further decremented for lower collinearity samples such as those represented by LOPOP. The Hoerl and Kennard (1970b) population again demonstrates the superiority of OLS at all sample sizes investigated.

Tables 13 through 15 present the relevant precision statistics for each coefficient for the HIPOP, HOPOP, and LOPOP matrices, respectively. Table 16 presents the differences between OLS and RR precision statistics pooled over sample sizes. It is interesting to note in these tables that RR produces a smaller bias in estimation for virtually all coefficients at all sample sizes for the HIPOP and LOPOP sets than does OLS despite the inclusion of k in equation (37) as a deliberate biasing factor. The exceptions among the 100 bias estimates are seven values found among the LOPOP matrices for samples of sizes 120 and 200. In these cases OLS and RR produce identical (to three places of accuracy)

Table 12

Ratio of Average LS_{cv} to Average $RIDGE_{cv}$[a]

	Population		
Sample Size	HIPOP	HOPOP	LOPOP
30	1.820	.919	1.182
60	3.194	.789	1.088
90	1.530	.706	1.053
120	1.460	.621	1.049
200	1.238	.532	1.032

Note. All entries are mean values for 25 samples averaged over 10 beta weights per sample.

[a] LS = ordinary least squares, RIDGE = ridge regression; cv = coefficient of variation.

Table 13

HIPOP Precision Statistics

Sample Size

		30				60				90				120				200			
Coefficient	Equation[a]	MSE	σ^2	BIAS^2	CV[b]	MSE	σ^2	BIAS^2	CV	MSE	σ^2	BIAS^2	CV	MSE	σ^2	BIAS^2	CV	MSE	σ^2	BIAS^2	CV
1	LS	.197	.078	.119	1327.13	.276	.109	.167	1572.63	.050	.020	.030	667.07	.018	.007	.011	399.26	.015	.006	.009	366.11
	R	.060	.023	.037	731.07	.031	.012	.019	524.94	.021	.008	.013	436.07	.008	.003	.006	273.73	.010	.003	.007	298.67
2	LS	.182	.069	.113	2.95	1.023	.380	.643	6.98	.066	.025	.041	1.77	.034	.014	.021	1.28	.018	.007	.011	.93
	R	.065	.025	.040	1.76	.025	.010	.015	1.08	.028	.011	.017	1.15	.015	.006	.009	.85	.009	.004	.006	.66
3	LS	.249	.099	.150	2.95	1.650	.625	1.026	7.60	.047	.019	.028	1.28	.031	.013	.019	1.05	.023	.009	.014	.90
	R	.030	.010	.020	1.02	.063	.019	.045	1.49	.019	.006	.013	.81	.012	.003	.009	.66	.013	.003	.009	.66
4	LS	.377	.144	.233	13.76	.403	.161	.242	14.25	.068	.026	.042	5.85	.054	.020	.034	5.21	.036	.014	.022	4.25
	R	.079	.028	.051	6.30	.047	.015	.032	4.86	.024	.008	.010	3.46	.021	.007	.015	3.26	.014	.005	.010	2.75
5	LS	.291	.112	.179	44.97	.144	.056	.088	31.66	.060	.024	.036	20.49	.028	.011	.017	14.05	.020	.008	.012	11.84
	R	.076	.030	.046	22.96	.029	.011	.018	14.28	.022	.008	.014	12.33	.010	.004	.006	8.26	.009	.003	.006	7.69
6	LS	.240	.092	.148	2.95	.347	.133	.214	3.54	.065	.026	.039	1.53	.046	.018	.028	1.28	.023	.009	.014	.91
	R	.055	.022	.033	1.41	.045	.018	.028	1.28	.022	.008	.014	.90	.018	.006	.012	.80	.010	.003	.007	.60
7	LS	.290	.114	.176	7.12	3.314	1.299	2.026	24.07	.108	.043	.065	4.35	.048	.019	.029	2.89	.026	.010	.016	2.13
	R	.060	.023	.036	3.23	.029	.012	.017	2.25	.028	.011	.017	2.19	.011	.004	.006	1.37	.007	.003	.004	1.09
8	LS	.133	.048	.085	1.46	.337	.129	.208	2.33	.057	.022	.035	.96	.035	.014	.021	.76	.015	.006	.009	.49
	R	.073	.018	.055	1.08	.056	.016	.040	.95	.035	.009	.009	.75	.023	.006	.017	.61	.015	.003	.012	.48
9	LS	.058	.022	.036	24.66	1.580	.588	.991	128.35	.028	.010	.018	17.09	.017	.006	.011	13.39	.007	.002	.004	8.28
	R	.035	.013	.022	19.00	.036	.011	.025	19.30	.016	.006	.011	13.03	.012	.004	.004	11.05	.005	.002	.003	7.20
10	LS	.102	.034	.068	7.21	3.150	1.190	1.964	40.14	.026	.010	.016	3.68	.022	.009	.013	3.33	.008	.003	.005	2.07
	R	.035	.014	.021	4.24	.018	.007	.011	2.99	.013	.005	.008	2.55	.011	.004	.004	2.42	.006	.002	.004	1.69
Averages	LS	.212	.081	.131	143.52	1.220	.466	.757	183.16	.058	.023	.035	72.41	.033	.013	.021	44.25	.019	.008	.012	39.79
	R	.057	.021	.036	79.21	.038	.013	.025	57.34	.023	.008	.015	47.33	.014	.005	.005	30.30	.010	.003	.007	32.15

Note: All entries are mean values based on 25 samples.

[a] LS = ordinary least squares; R = ridge regression.

[b] CV = coefficient of variation.

Table 14

HOPOP Precision Statistics

Sample Size

		30				60				90				120				200			
Coefficient	Equation[a]	MSE	σ^2	BIAS^2	CV	MSE	σ^2	BIAS^2	CV	MSE	σ^2	BIAS^2	CV	MSE	σ^2	BIAS^2	CV	MSE	σ^2	BIAS^2	CV
1	LS	.045	.017	.028	1.19	.034	.014	.021	1.04	.020	.008	.013	.81	.013	.005	.008	.65	.006	.002	.004	.44
	R	.060	.006	.054	1.38	.034	.005	.029	1.04	.035	.003	.032	1.06	.031	.002	.030	1.00	.025	.002	.023	.89
2	LS	.021	.008	.012	.63	.016	.006	.009	.55	.009	.004	.005	.42	.004	.001	.002	.26	.003	.001	.002	.24
	R	.034	.004	.029	.80	.026	.005	.021	.70	.019	.002	.017	.61	.014	.001	.013	.51	.011	.001	.010	.46
3	LS	.048	.019	.029	.59	.012	.005	.007	.29	.017	.007	.010	.35	.007	.003	.004	.23	.004	.002	.002	.17
	R	.043	.010	.033	.56	.017	.004	.013	.35	.019	.004	.014	.37	.012	.003	.010	.29	.008	.002	.007	.24
4	LS	.022	.009	.013	1.35	.007	.003	.004	.75	.003	.001	.002	.45	.003	.001	.002	.49	.003	.001	.002	.46
	R	.020	.006	.014	1.27	.010	.002	.008	.92	.006	.001	.005	.70	.005	.001	.004	.66	.005	.001	.004	.62
5	LS	.129	.052	.077	.78	.041	.015	.026	.44	.024	.009	.015	.34	.010	.003	.006	.21	.011	.004	.007	.22
	R	.292	.010	.283	1.17	.236	.009	.227	1.05	.190	.005	.184	.94	.159	.003	.156	.86	.136	.004	.131	.80
6	LS	.132	.053	.080	.45	.053	.021	.032	.28	.038	.014	.024	.24	.023	.008	.015	.19	.012	.004	.008	.14
	R	.387	.006	.381	.77	.301	.008	.293	.68	.268	.005	.262	.64	.238	.003	.235	.60	.201	.003	.197	.55
7	LS	.032	.012	.020	.61	.011	.005	.007	.36	.015	.006	.009	.41	.008	.003	.005	.31	.004	.002	.003	.22
	R	.119	.005	.113	1.18	.066	.005	.061	.88	.055	.004	.052	.80	.047	.002	.045	.74	.038	.002	.036	.67
8	LS	.023	.008	.015	.39	.015	.006	.010	.32	.015	.005	.010	.31	.010	.004	.006	.26	.005	.002	.003	.17
	R	.044	.006	.039	.54	.029	.005	.024	.43	.025	.004	.022	.41	.018	.002	.016	.34	.011	.001	.010	.27
9	LS	.008	.003	.005	1.09	.010	.004	.006	1.23	.007	.003	.004	.99	.004	.002	.003	.79	.002	.001	.001	.48
	R	.007	.002	.005	1.00	.010	.002	.008	1.23	.007	.002	.005	1.03	.006	.002	.004	.94	.003	.001	.002	.62
10	LS	.036	.012	.024	2.01	.013	.004	.009	1.22	.007	.003	.004	.88	.006	.002	.003	.79	.002	.001	.001	.48
	R	.013	.005	.008	1.23	.008	.003	.005	.94	.006	.002	.004	.82	.005	.002	.003	.77	.003	.001	.002	.57
Averages	LS	.049	.019	.030	.91	.021	.008	.013	.65	.015	.006	.010	.52	.009	.003	.006	.42	.005	.002	.003	.30
	R	.006	.006	.096	.99	.074	.005	.069	.82	.063	.003	.060	.74	.054	.002	.052	.67	.044	.002	.042	.57

Note: All entries are mean values based on 25 samples.

[a] LS = ordinary least squares; R = ridge regression.

[b] CV = coefficient of variation.

Table 15

LOPOP Precision Statistics

		Sample Size																			
		30				60				90				120				200			
Coefficient	Equation[a]	MSE	σ^2	BIAS^2	CV	MSE	σ^2	BIAS^2	CV	MSE	σ^2	BIAS^2	CV	MSE	σ^2	BIAS^2	CV	MSE	σ^2	BIAS^2	CV
1	LS	.194	.074	.120	8.81	.073	.029	.044	5.40	.038	.015	.023	3.88	.015	.006	.010	2.47	.011	.004	.006	2.05
	R	.135	.052	.083	7.34	.061	.024	.037	4.94	.033	.013	.020	3.65	.014	.005	.009	2.37	.010	.004	.006	1.97
2	LS	.127	.050	.077	5.42	.064	.025	.038	3.83	.044	.017	.027	3.18	.020	.008	.012	2.17	.016	.006	.010	1.92
	R	.091	.036	.055	4.59	.052	.021	.031	3.46	.039	.015	.024	3.01	.018	.007	.011	2.06	.015	.006	.009	1.85
3	LS	.179	.071	.108	2.07	.090	.036	.054	1.46	.036	.014	.022	.92	.031	.012	.019	.85	.017	.007	.010	.63
	R	.087	.035	.042	1.44	.067	.026	.041	1.27	.030	.012	.018	.84	.025	.010	.015	.77	.015	.006	.009	.59
4	LS	.273	.101	.171	3.82	.112	.040	.072	2.45	.039	.015	.025	1.45	.034	.013	.021	1.35	.028	.011	.017	1.21
	R	.175	.061	.114	3.06	.098	.032	.066	2.29	.037	.013	.024	1.41	.033	.012	.021	1.32	.027	.010	.017	1.20
5	LS	.247	.092	.155	4.53	.076	.030	.045	2.51	.046	.018	.028	1.96	.020	.008	.012	1.28	.015	.006	.009	1.12
	R	.172	.064	.108	3.79	.060	.024	.036	2.24	.040	.016	.024	1.83	.018	.007	.011	1.21	.014	.005	.009	1.08
6	LS	.173	.068	.105	4.47	.067	.027	.040	2.78	.041	.016	.025	2.17	.026	.010	.016	1.71	.016	.006	.010	1.35
	R	.122	.048	.074	3.75	.054	.021	.033	2.50	.037	.014	.023	2.06	.024	.009	.015	1.65	.015	.006	.009	1.31
7	LS	.127	.049	.079	6.48	.045	.017	.027	3.84	.042	.017	.025	3.72	.024	.009	.014	2.79	.012	.005	.007	2.01
	R	.088	.034	.054	5.40	.037	.015	.023	3.52	.038	.015	.023	3.54	.022	.009	.013	2.66	.012	.005	.007	1.95
8	LS	.078	.030	.047	11.95	.041	.016	.025	8.67	.039	.015	.024	8.52	.021	.008	.013	6.24	.010	.004	.006	4.20
	R	.063	.024	.039	10.75	.037	.015	.022	8.21	.036	.014	.022	8.12	.019	.008	.012	5.94	.009	.005	.006	4.10
9	LS	.049	.019	.030	3.09	.050	.017	.033	3.14	.031	.011	.020	2.46	.016	.005	.011	1.79	.007	.003	.005	1.21
	R	.038	.015	.023	2.75	.044	.015	.029	2.95	.028	.010	.018	2.35	.015	.005	.010	1.72	.007	.003	.004	1.17
10	LS	.143	.048	.095	2.27	.063	.021	.042	1.50	.031	.013	.022	1.12	.032	.012	.019	1.06	.010	.004	.006	.61
	R	.099	.032	.067	1.89	.051	.017	.034	1.36	.032	.012	.020	1.07	.028	.011	.017	1.01	.009	.004	.006	.58
Averages	LS	.159	.060	.099	5.29	.068	.026	.042	3.56	.039	.015	.024	2.94	.024	.009	.015	2.17	.014	.006	.009	1.63
	R	.107	.040	.067	4.47	.056	.021	.035	3.27	.035	.013	.022	2.79	.022	.008	.013	2.07	.013	.005	.008	1.58

Note: All entries are mean values based on 25 samples.

[a] LS = ordinary least squares; R = ridge regression.

[b] CV = coefficient of variation.

Table 16

Mean Differences of Precision Statistics

Between Least Squares and RIDGE

	Population											
	HIPOP				HOPOP				LOPOP			
Coefficient	MSE	σ^2	$BIAS^2$	CV^a	MSE	σ^2	$BIAS^2$	CV	MSE	σ^2	$BIAS^2$	CV
1	.085	.034	.051	413.54	-.013	.006	-.019	-.25	.016	.006	.010	.47
2	.236	.088	.148	1.68	-.010	.001	-.012	-.20	.011	.004	.007	.31
3	.373	.145	.228	1.83	-.002	.003	-.005	-.04	.026	.010	.018	.20
4	.151	.060	.091	4.54	-.002	.001	-.002	-.13	.023	.010	.013	.20
5	.079	.031	.048	11.50	-.160	.010	-.170	-.57	.020	.008	.012	.25
6	.114	.044	.070	1.04	-.227	.015	-.242	-.39	.014	.006	.008	.24
7	.730	.286	.446	6.09	-.051	.002	-.053	-.47	.011	.004	.006	.35
8	.074	.033	.045	.43	-.012	.001	-.013	-.11	.005	.001	.003	.49
9	.317	.118	.199	24.44	.000	.001	-.001	-.05	.004	.001	.003	.15
10	.645	.243	.404	8.51	.006	.002	.004	.21	.012	.004	.008	.13

Note: All entries are mean values based on the difference between sets of 25 samples pooled over sample sizes.

[a] CV = coefficient of variation.

bias values. In Table 14 the inappropriateness of RR for the HOPOP based samples is again evident. For most coefficients the variance of the estimate has been reduced as expected but the bias has grown large enough to offset this improvement and thus the MSE is worse, in general, for RR than for OLS. This information is summarized in Table 16 where negative values indicate the superior accuracy (i.e., smaller variance, less bias, smaller MSE, etc.) of OLS.

CHAPTER V

SUMMARY

The ridge regression technique advocated by Hoerl and Kennard (1970a, 1970b) for use in regression analyses when high collinearity is present among the predictors was evaluated by comparing it with the ordinary least squares and unit weighting methods of combining predictors so as to form a linear composite. It is concluded that the RR technique is preferable to OLS in certain restricted situations. Specifically, if one's intent is to optimize the prediction of a criterion by the composite, unit weighting with signs for the predictor variables determined by their zero-order validity coefficients is probably preferable to the calculation of a ridge regression. When sample size is large enough (for instance, greater than 200, Schmidt, 1971) so that OLS weights are expected to yield better prediction in subsequent samples than will unit weights, the indication from the present study is that RR will be very little, if at all, better. Thus, for large samples OLS seems to provide weights which are not significantly inferior to RR weights for prediction.

If the purpose of the regression analysis is structural interpretation however, RR, based on the empirical populations evaluated above, yields substantially more accurate weight estimates than does

OLS for certain types of matrices. Ridge regression, on the average, exhibited less bias, smaller variance, and a smaller overall MSE than OLS for all coefficients estimated in samples drawn from two empirical populations: HIPOP, in which the pairwise intercorrelation was high and consistent throughout the matrix, and LOPOP, in which pairwise values were low but also consistent across 10 predictors.

The ridge technique was dominated by OLS from both a predictive and structural interpretation perspective when evaluation was based on samples drawn from the HOPOP matrix. This matrix exhibited a number of characteristics which are uncommon in social science data sets. The wide dispersion of pairwise intercorrelations and the presence of both high positive and high negative validities contributed to the formation of a population beta vector with large absolute values. This would appear to be the prime reason for ridge regression's failure to yield better precision statistics than OLS in samples from the HOPOP matrix. RR functions in general to reduce the absolute value of the beta estimates in the sample which are considered inflated due to sample specific error and high multicollinearity. However, in the Hoerl and Kennard (1970b) population the collinearity is not consistent throughout the matrix and the validities for some relatively independent predictors are quite high (see Table 3). Applying RR to samples from a population with these characteristics will, it appears, result in overcorrection for the properly large weights and undercorrection for the others.

It is concluded therefore that RR can be used to more precisely estimate beta weights even when an analytic determination of the bias parameter is not available. The prime consideration in the decision to use RR over OLS is the type of matrix one has to analyze. It is apparent that the determinant is insufficient as an indicator of the degree of collinearity as it is not sensitive to the distribution of pairwise correlations in the matrix. Two suggestions may be made on the basis of the present study. First, RR is a preferable method of analysis to the extent that the correlation matrix is unidimensional. This can be evaluated by considering the eigenvalues and eigenvectors for the matrix. Secondly, the diagonal values of the correlation matrix's inverse should not be widely dispersed. In effect, this means that any p-1 combination of the predictors should predict the p^{th} predictor equally well. This rule is an indication of the consistency of the intercorrelation level in the matrix.

The ridge regression technique has been criticized justifiably as being an arbitrary method when applied to any one sample matrix which requires excessive subjective interpretation and final identification of an adequate result (Conniffe & Stone, 1973; Smith, 1976). The element of subjectivity, whether in defining an arbitrary bias (k) selection rule or in visually examining the ridge trace for a stable solution, cannot be denied at the moment. However, it is concluded on the basis of this study that reasonable decision rules can be established which, for appropriate data sets, will function better than ordinary least squares. Thus, the author is in agreement with Churchill (1975), Hoerl and Kennard (1970b), and Smith and Goldstein (1975) in concluding that RR does offer potential improvements over OLS. It is hoped that an analytic derivation of an optimal bias parameter will be developed as well as objective methods of determining the appropriateness of any particular data set for the ridge technique.

BIBLIOGRAPHY

BIBLIOGRAPHY

Anderson, H. E., Jr., & Fruchter, B. Some multiple correlation and predictor selection methods. *Psychometrika*, 1960, *25*, 59-76.

Bartlett, M. S. Tests of significance in factor analysis. *British Journal of Psychology*, *Statistical Section*, 1950, *3*, 77-85.

Berkson, J. Are there two regressions? *Journal of the American Statistical Association*, 1950, *45*, 164-180.

Binder, A. Considerations of the place of assumptions in correlational analysis. *American Psychologist*, 1959, *14*, 504-510.

Blum, M., & Naylor, J. C. *Industrial psychology: Its theoretical and social foundations*. Third edition. New York: Harper and Row, 1968.

Burket, G. A. A study of reduced rank models for multiple prediction. *Psychometric Monographs*, 1964, No. 12.

Cattell, R. B. The scree test for the number of factors. *Multivariate Behavioral Research*, 1966, *1*, 245-251.

Churchill, G. A., Jr. A regression estimation method for collinear predictors. *Decision Sciences*, 1975, *6*, 670-687.

Claudy, J. G. A comparison of five variable weighting procedures. *Educational and Psychological Measurement*, 1972, *32*, 311-322.

Cochrane, C., & Orcutt, G. H. Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, 1949, *44*, 32-61.

Cohen, J. Multiple regression as a general data analytic system. *Psychological Bulletin*, 1968, *70*, 426-443.

Cohen, J., & Cohen, P. *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, N.J.: Erlbaum Associates, 1975.

Coniffe, D., & Stone, J. A critical view of ridge regression. *The Statistican*, 1973, *22*, 181-187.

Cooley, W. W., & Lohnes, P. R. Multivariate Data Analysis. New York: Wiley, 1971.

Cureton, E. E. Validity. In E. F. Lindquist (Ed.), Educational Measurement. Washington: American Council on Education, 1950.

Cureton, E. E. Multivariate Psychological Statistics. Unpublished manuscript, The University of Tennessee, Knoxville, 1962.

Cureton, E. E. A measure of the average intercorrelation. Educational and Psychological Measurement, 1971, 31, 627-628.

Darlington, R. B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69, 161-182.

Dawes, R. M., & Corrigan, B. Linear models in decisin making. Psychological Bulletin, 1974, 81, 95-106.

Draper N., & Smith, H. Applied regression analysis. New York: Wiley, 1966.

Durbin, J., & Watson, G. S. Testing for serial correlation in least squares regression. Biometrika, 1950, 37, 409-428.

Einhorn, H. J., & Hogarth, R. M. Unit weighting schemes for decision making. Organizational Behavior and Human Performance, 1975, 13, 171-192.

Farrar, D. E., & Glauber, R. R. Multicollinearity in regression analysis: The problem revisited. Review of Economics and Statistics, 1967, 49, 92-107.

Finn, J. D. A general model for multivariate analysis. New York: Holt, Rinehart, and Winston, Inc., 1974.

Furnival, G. M., & Wilson, R. W., Jr. Regressions by leaps and bounds. Technometrics, 1974, 16, 499-511.

Geary, R. C. Non-linear functional relationship between two variables when one variable is controlled. Journal of the American Statistical Association, 1953, 48, 94-103.

Gordon, R. A. Issues in multiple regression. American Journal of Sociology, 1967, 73, 591-616.

Gorman, J. W., & Toman, R. J. Selection of variables for fitting equations to data. Technometrics, 1966, 8, 27-51.

Guilkey, D. K., & Murphy, J. L. Directed ridge regression techniques in cases of multicollinearity. Journal of the American Statistical Association, 1975, 70, 769-775.

Guttman, L. To what extent can communalities reduce rank? Psychometrika, 1958, 23, 297-309.

Herzberg, P. A. The parameters of cross-validation. Psychometric Monographs, 1969, No. 14.

Hocking, R. R., & Leslie, R. N. Selection of the best subset in regression analysis. Technometrics, 1967, 9, 531-540.

Hoerl, A. E. Application of ridge analysis to regression problems. Chemical Engineering Progress, 1962, 58, 54-59.

Hoerl, A. E., & Kennard, R. W. Ridge regression: Biased estimation for non-orthogonal problems. Technometrics, 1970a, 12, 55-67.

Hoerl, A. E., & Kennard, R. W. Ridge regression: Applications to nonorthogonal problems. Technometrics, 1970b, 12, 69-82.

Hoerl, A. E., Kennard, R. W., & Baldwin, K. F. Ridge regression: Some simulations. Communications in Statistics, 1975, 4, 105-123.

Horst, P. The prediction of personal adjustment. New York: Social Science Research Council, 1941.

Horst, P., & MacEwan, C. Predictor-elimination techniques for determining multiple prediction batteries. Psychological Reports, 1960, 7, 19-50.

Jeffers, J. N. R. Two case studies in the application of principal component analysis. Applied Statistics, 1967, 16, 225-236.

Johnston, J. Econometric methods. Second edition. New York: McGraw-Hill, Inc., 1972.

Joliffe, I. T. Discarding variables in a principal component analysis. 1. Artifactual data. Applied Statistics, 1973, 22, 21-31.

Kaiser, H. F. A measure of the average intercorrelation. Educational and Psychological Measurement, 1968, 28, 245-247.

Klein, L. R. An introduction to econometrics. New Jersey: Prentice-Hall, 1962.

Klein, L. R., & Nakamura, M. Singularity in the equation systems of econometrics: Some aspects of the problem of multicollinearity. International Economic Review, 1962, 3, 274-299.

Klein, S. P., & Evans, F. R. An examination of the validity of nine experimental tests for predicting success in law school. Educational and Psychological Measurement, 1968, 28, 909-913.

Klein, S. P., & Evans, F. R. Early prediction of independent accomplishment. Unpublished manuscript, 1969.

Kmenta, J. Elements of econometrics. New York: The MacMillan Co., 1971.

LaMotte, L. R., & Hocking, R. R. Computational efficiency in the selection of regression variables. Technometrics, 1970, 12, 83-93.

Lawless, J. F., & Wang, P. A simulation study of ridge and other regression estimates. Communications in Statistics, 1976, 5, 307-323.

Lawshe, C. H., & Shucker, R. E. The relative efficiency of four test weighting methods in multiple prediction. Educational and Psychological Measurement, 1959, 19, 103-114.

Li, J. C. R. Statistical inference II. Ann Arbor, MI: Edwards Brothers, Inc., 1974.

Lord, F. M. Efficiency of prediction when a regression equation from one sample is used in a new sample. Research Bulletin 50-40. Princeton, N.J.: Educational Testing Service, 1950.

Marquardt, D. W. Generalized inverses, ridge regression, biased linear estimation and non-linear estimation. Technometrics, 1970, 12, 591-612.

Marquardt, D. W., & Snee, R. D. Ridge regression in practice. The American Statistician, 1975, 29, 3-20.

Mason, R. L., Gunst, R. F., & Webster, J. T. Regression analysis and problems of multicollinearity. Communications in Statistics, 1975, 4, 277-292.

Massy, W. F. Principal components regression in exploratory statistical research. Journal of the American Statistical Association, 1965, 60, 234-256.

McDonald, G. C., & Galarneau, D. I. Monte-Carlo evaluation of some ridge-type estimators. Journal of the American Statistical Association, 1975, 70, 407-416.

McDonald, G. C., & Schwing, R. C. Instabilities of regression estimates relating air pollution to mortality. Technometrics, 1972, 15, 463-481.

Meyer, E. P. A measure of the average intercorrelation. Educational and Psychological Measurement, 1975, 35, 67-72.

Neter, J., & Wasserman, W. Applied linear statistical models. Homewood, Ill.: Irwin, 1974.

Newhouse, J. P., & Oman, S. D. An evaluation of ridge estimators. U.S. Air Force Project Rand R-7161PR, 1971.

Nicholson, G. E. Prediction in future samples. In I. Olkin, et al. (Eds.), Contributions to probability and statistics. Palo Alto, CA: Stanford, 1960, pp. 322-330.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. Statistical package for the social sciences. New York: McGraw-Hill, 1975.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Overall, J. E., & Klett, C. J. Applied multivariate analysis. New York: McGraw-Hill, 1972.

Rao, P., & Miller, R. L. Applied econometrics. Belmont, CA: Wadsworth Publishing Co., Inc., 1971.

Rock, D. A., Linn, R. L., Evans, F. R., & Patrick, C. A comparison of predictor selection techniques using Monte Carlo methods. Educational and Psychological Measurement, 1970, 30, 873-884.

Roose, J. E., & Doherty, M. E. Judgment theory applied to the selection of life insurance salesmen. Organizational Behavior and Human Performance, 1976, 16, 231-249.

Sastry, M. V. R. Some limits in the theory of multicollinearity. The American Statistician, 1970, 24, 39-40.

Scheffe, H. The analysis of variance. New York: Wiley, 1959.

Schmidt, F. L. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. Educational and Psychological Measurement, 1971, 31, 699-714.

Schmitt, N., & Coyle, B. W. Applicant decisions in the employment interview. Journal of Applied Psychology, 1976, 61, 184-192.

Schmitt, N., Coyle, B. W., & Rauschenberger, J. A Monte Carlo evaluation of three formula estimates of cross-validated multiple correlation. Psychological Bulletin, 1977, in press.

Sevier, F. A. C. Testing the assumptions underlying multiple regression. Journal of Experimental Education, 1957, 25, 323-330.

Silvey, S. D. Multicollinearity and imprecise estimation. Journal of the Royal Statistical Society, Series B, 1969, 31, 539-552.

Smith, A. F. M., & Goldstein, M. Ridge regression: Some comments on a paper of Connife & Stone. The Statistician, 1975, 24, 61-66.

Smith, V. K. A note on ridge regression. Decision Sciences, 1976, 7, 562-566.

Snee, R. D. Some aspects of non-orthogonal data analysis, part I: Developing prediction equations. Journal of Quality Technology, 1973, 5, 67-79.

Swindel, B. F. Instabilities of regression coefficients illustrated. The American Statistician, 1974, 28, 63-65.

Tatsuoka, M. M. Multivariate analysis: Techniques for educational and psychological research. New York: Wiley, 1971.

Trattner, M. H. Comparison of three methods for assembling aptitude test batteries. Personnel Psychology, 1963, 16, 221-232.

Tukey, J. W. One degree of freedom for nonadditivity. Biometrics, 1949, 5, 232-242.

Tupes, E. C. A note on "validity and nonlinear heteroscedastic models." Personnel Psychology, 1964, 17, 59-61.

Vinod, H. D. Application of new ridge regression methods to a study of Bell System scale economics. Journal of the American Statistical Association, 1976, 71, 835-841.

Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. Psychological Bulletin, 1976, 83, 213-217.

Wainer, H., & Thissen, D. Three steps towards robust regression. Psychometrika, 1976, 41, 9-34.

Wampler, R. H. A report on the accuracy of some widely used least squares computer programs. Journal of the American Statistical Association, 1970, 65, 549-565.

Webster, J. T., Gunst, R. F., & Mason, R. L. Latent root regression analysis. Technometrics, 1974, 16, 513-522.

Wesman, A. G., & Bennett, G. K. Multiple regression vs. simple addition of scores in prediction of college grades. Educational and Psychological Measurement, 1959, 19, 243-246.

Wherry, R. J., Sr. A new formula for predicting the shrinkage of the coefficient of multiple correlation. The Annals of Mathematical Statistics, 1931, 2, 440-457.

Wherry, R. J., Sr. Underprediction from overfitting: 45 years of shrinkage. Personnel Psychology, 1975, 28, 1-18.

Wherry, R. J., Sr., Naylor, J. C., Wherry, R. J., Jr., & Fallis, R. F. Generating multiple samples of multivariate data with arbitrary population parameters. Psychometrika, 1965, 30, 303-313.

Wilks, S. S. Certain generalizations in analysis of variance. *Biometrika*, 1932, *24*, 471-494.

Winer, B. J. *Statistical principles in experimental design*. Second edition. New York: McGraw-Hill, 1971.

Wishart, J. The mean and second moment of the multiple correlation coefficient in samples from a normal population. *Biometrika*, 1931, *22*, 353-361.