MODEL ESTIMATION USING RIDGE REGRESSION WITH THE VARIANCE NORMALIZATION CRITERION

WAN-FUNG LEE

MODEL ESTIMATION USING RIDGE REGRESSION WITH THE VARIANCE NORMALIZATION CRITERION

by

WAN-FUNG LEE, B.Sc., M.Sc. ©

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Education
Department of Educational Foundations
Memorial University of Newfoundland

February 1979

ABSTRACT

Structural equation model building has been extensively used in the social sciences. The ordinary least squares (OLS) regression technique has been the standard technique used in the single equation method of estimation. OLS regression estimates are erroneous, however, due to the presence of multicollinearity which is attributable to an absence of control over the survey data and to an intrinsic property of structural equation models. The inadequacy of the OLS regression technique when applied to ill-conditioned data was discussed in chapter II.

Ridge regression, developed by Hoerl and Kennard (1970), is the most promising technique for coping with the multicollinearity problem. However, the technique is inadmissible due to the stochasticity of the estimation criterion. An exposition of ridge regression theory was given in chapter III. In chapter IV, the dilemma of ridge regression was analyzed and a new criterion, called the variance normalization criterion was developed. With this criterion all the difficulties encountered by Hoerl and Kennard's version of ridge regression are avoided.

In chapter V, simple ridge regression with the variance normalization criterion was applied to a 5-stage human capital problem which used the Malmö data. Through this example and through the theoretical arguments discussed in chapter II, III, and IV, the following goals of the study were achieved: (1) the superiority of simple ridge regression over ordinary least square regression was demonstrated; (2) Hoerl and Kennard's

version of ridge regression was modified in order to achieve more satisfactory results; and (3) it was demonstrated that simple ridge regression with the variance normalization criterion is superior both to ridge regression estimation procedures using the mean square error criterion and the OLS procedure.

ACKNOWLEDGEMENTS

The author is deeply indebted to Dr. Wo-Shun Luk for his assistance in this study. His computer program, written especially for this research, has made it possible for the author to compare various ridge regression procedures, and thereby gain insight into the characteristics of the theory developed in this study. The author is also obliged to Dr. Charles Lee for his constructive criticism, suggestion, and correction, and to Mr. Donold Heale for his assistance in the computer programing in the early stage of this study.

The author's deepest gratitude is to Professor Jeffrey Bulcock for his continuous assistance, encouragement and support on this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION TO THE STUDY

Statement of the Problem

Structural equation model building, also known as causal modelling or path analysis has been extensively used in sociological research in the past decade. The basic technique used in this type of model building is ordinary least squares (OLS) regression. In fact, ordinary least squares regression has been so extensively used in all the sciences, that it can be described as a standard statistical technique (McCabe, 1978).

The least squares technique is theoretically sound; however, empirically it is inadmissible. One of the most important empirical difficulties is the multicollinearity problem which can be defined as the departure of the predictor vectors "X's" from orthogonality or, equivalently the existence of linearity, among the explanatory variables. In sociological research, due to the lack of researcher control over the variables, the use of sets of nearly independent predictor variables is virtually unknown; thus, multicollinearity is an unavoidable problem.

In structural equation model building, the multicollinearity problem is twofold: first, as has been mentioned, multicollinearity is unavoidable due to the lack of control over the explanatory variables; and second, multicollinearity is an intrinsic property of structural equation models due to the fact that in multi-stage models of the path variety the better the specification at one stage the less the likelihood of linear independence at subsequent stages. Therefore, it is important that, the analyst be aware of the existence of the problem and be familiar with the technique for coping, or ideally resolving, the multicollinearity problem. The most promising technique for reducing the harmful effects of multicollinearity is called "ridge regression" (RR) which was developed in 1970 by Hoerl and Kennard (1970a, 1970b). The ridge regression procedure involves augmenting the diagonal of the normal equation matrix with a small positive quantity "k" in order to produce regression coefficients with smaller variance at the expense of introducing a small bias. The main difference between OLS regression and ridge regression lies in the criterion of goodness of estimation. In ridge regression, the criterion is the minimum total mean squared errors (MSE) as opposed to the minimum sum of squared errors (SSE) used in ordinary least squares.

The main difficulty with ridge regression has always been the estimation of the optimal k which produces the minimum total mean squared errors. Theoretically, for any regression problem there always exists an optimal k; however, it depends on two unknown parameters, the error variance, σ^2, and the true regression coefficient vector β. These two parameters are population constants not universal constants; due to this nature of "k", its optimum magnitude cannot feasibly be estimated. Thus, although most researchers in this field have been pursuing methods of estimating an optimal k which gives a minimum MSE, and though more than fifteen methods have been suggested, none can be considered satisfactory. The purpose of this research is not to add another method to the extant methods of "solving" this virtually unsolvable problem. Instead, as the substantive dimension of the research, the minimum MSE criterion is abandoned in favour of a "weaker" unifunctional criterion as an alternative in order to achieve more satisfactory empirical results. Further, it is also the purpose of this research to demonstrate the superiority of ridge regression over OLS regression both through theoretical argument and through the application or ridge regression to model building research in education.

Significance of the study

It is well known that the ordinary least squares regression procedure does not produce satisfactory results (Stein, 1955), especially when the data set has high degree of multicollinearity (Hoerl & Kennard, 1970a, Marquardt, 1970). It is also well known that multicollinearity is always present in any multiple regression problem, and that its seriousness is a matter of degree (Kmenta, 1971). It has also been pointed out in the previous section that multicollinearity is an intrinsic problem of the multistage structural equation model: the better specified the model, the higher its degree of the multicollinearity. The ill-effect of the problem in the general case will produce estimates with large variances and this in turn leads to estimates that are unreliable (even of wrong sign) and sensitive to sampling error or model misspecification (Hoerl and Kennard 1970a). It is therefore of the utmost importance for the statistical analyst to realize the existence and seriousness of this problem before making any statistical inference or claim.

Among all alternatives to OLS (Dempster et.al. 1977, have studied 56 of them) it has been generally established that ridge regression, which is designed to cope with the problem of multicollinearity, is the best and most promising one, as compared to those alternatives currently under study, such as shrunken estimators, and principal component estimators. Like all alternatives to OLS, the ridge estimator is a biased one; however, it generally has much smaller variance and therefore is less sensitive to sampling fluctuation or model misspecification and possesses more accurate predicting power.

Due to the seriousness of the inadequacy of OLS regression, research in ridge regression is of interest to all sciences, social, natural, or applied. Further, the modification developed in this study would render the application of ridge regression appropriate, indeed necessary, to any multiple regression problem in the interests of scientific parsimony and accuracy, not just to those problems with a high degree of multicollinearity as was originally intended.

To summarize, this study is of utmost importance since (i) the widely used regression technique, that is OLS regression, is inadmissable, (ii) ridge regression is the most promising alternative to OLS regression, (iii) the modification of Hoerl and Kennard's ridge regression would widen and even replace OLS regression, (iv) the theoretical results are of interest to researchers in the natural and social sciences, both pure and applied.

Limitation of the study

Ridge regression developed by Hoerl and Kennard in 1970, is of two types, simple ridge regression (SRR) and generalized ridge regression (GRR). Simple ridge regression is a procedure which involves augmenting all the diagonal elements of the normal equation matrix with the same constant, while in generalized ridge regression the diagonal elements of the canonical normal equation matrix are augmented with different constants. Hocking, Speed and Lynn (1976) have proved that theoretically generalized ridge regression is superior to simple ridge regression; however, empirically it could be just the opposite. In this study, the research is limited to simple ridge regression. The modification of generalized ridge regression in the same manner as proposed here for simple ridge regression will be pursued in further research.

The modification of simple ridge regression suggested in this study requires the replacement of the minimum MSE criterion with a weaker criterion, called the "variance normalization criterion", which is justified by its single purpose - the minimization of the effect of multicollinearity through normalizing the variance inflation factor. In this way most of the estimation dilemma encountered by ridge regression using the minimum MSE criterion becomes avoidable. The normalization criterion is not a novel idea, it is a development stemming from Marquardt's rule of thumb (Marquardt, 1970). This modified ridge regression has not been examined using Monte Carlo simulation methods. Although this is desirable, it is beyond the scope of present study.

Due to the nature of the present research, the application of ridge regression is limited to educational examples, which often have a much lower degree of multicollinearity than many problems in engineering or economics.

In short, this study is limited to simple ridge regression and to educational examples. It lacks rigorous examination by using a Monte Carlo simulation method, and it is only the first stage of a series of research studies designed to refine, to generalize and to extend through application, the theory of ridge regression analysis.

CHAPTER II

THEORY (I)

Multicollinearity and its Effects on OLS Regression Procedure

The classical multiple regression model is given by:

$$y = X\beta + \varepsilon \qquad (2.1)$$

Where y is an (n x 1) vector of observations on the dependent variable, X is an (n x p) matrix of observations on p explanatory variables; where for convenience X and y are scaled so that X'X and X'y give the correlation matrices; and where β is the (k x 1) vector of regression coefficients and ε is an (n x 1) vector of the random error of y.

In ordinary least squares estimation it is assumed

(1) that the X's are nonstochastic;[1]

(2) that there are no (exact) linear relationships between the explanatory variables; i.e., X is of full rank; and

(3) that the error terms have independent normal distribution with zero mean and constant variance, that is $\varepsilon \sim NID(o,\sigma^2)$

The robustness, and the effects of the violation of these assumptions have been thoroughly discussed by many researchers in more advanced texts. Here we are interested in the second assumption. Mathematically, insofar as there is no exact linear

[1]This assumption can be relaxed to accommodate stochastic variables (see Johnston, 1972, ch. 9, pp. 267).

relationship between any two predictors, the model can be solved. However, from a practical point of view this is not enough. If a nearly linear relationship exists among the predictors, the "solution" will probably prove unacceptable, in that such a "solution" will be unstable, unreliable, and hard, if not impossible, to interpret. This problem is generally called the multicollinearity problem.

In practice, and especially in the social sciences, most predictor variable sets are intercorrelated; thus, the problem severity is a matter of degree. As pointed out by Kmenta (1971), multicollinearity is a matter of degree not of kind.

The OLS solution of the regression model can be written as:

$$\hat{\beta} = (X'X)^{-1} X'y \tag{2.2}$$

with the variance - covariance matrix,

$$\text{cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}. \tag{2.3}$$

When there is a high level of multicollinearity, the determinant $|X'X|$ will be very small and $(X'X)^{-1}$ will have very large entries (in the extreme case, $|X'X| = 0$ and $(X'X)^{-1}$ would not exist), therefore, each estimated coefficient would have a very large variance and covariance, which is the case for all the problems associated with multicollinearity as will be discussed below.

There are several ways of measuring the severity of multicollinearity. Marquardt and Snee (1975) have pointed out that, the maximum variance inflation factor is the best single measure. The variance inflation factors (VIF) are the diagonal elements of the inverse of the simple correlation matrix. The variance inflation factor of each term is a measure of the collective impact of the interdependency of the explanatory variables on the variance of the regression coefficient of that term; and its value gives us an indication of the number of times the variance has been inflated. When maximum VIF is used to measure the level of multicollinearity, we have 1 for no collinearity and infinity for perfect collinearity. The "degree" of multicollinearity can, however, be transformed into a scale of zero to one by defining the degree of multicollinearity "D" as:

$$D = D_{max} = \frac{2}{\pi}\tan^{-1}(V_{max} - 1) \tag{2.4}$$

where we have used V_{max} for the maximum variance inflation factor, and have $D = 0$ for no collinearity and $D = 1$ for perfect collinearity.[2] With this scale, $D \geq 0.7$ may be generally considered serious. Collinearity of this magnitude is common in the social sciences.

The claim that $D \geq 0.7$ is generally serious is rather subjective. Whether the degree of multicollinearity is serious or not sometimes depends on the size of the true regression coefficients of the problem; even for $D = 0.5$ (equivalent to $V_{max} = 2$) or smaller, if one of the regression coefficients is small enough such that the inflated variance spans zero, the estimated OLS regression coefficient can be of the wrong sign. In this case the effect of multicollinearity is definitely

[2] $\tan^{-1}(V_{max} - 1)$ is measured in radians

serious. Therefore, the D value magnitude indicative of the extent to which the problem of multicollinearity may be considered serious depends on the case. Generally, when $D = 0.7$, the maximum variance inflation factor is about 3 and the relation between V_{max} and D becomes steeper. In these instances D values ≥ 0.7 can be considered as serious regardless of the size of the regression coefficients.

In view of the fact that the maximum effect of multicollinearity is manifest in factor space when the model is expressed in its canonical form, it may be best to measure the absolute degree of multicollinearity in factor space, for which $V_{max} = \frac{1}{\lambda min}$ and

$$D \leq D_c = \frac{2}{\pi}\tan^{-1}\left(\frac{1}{\lambda min} - 1\right) \qquad (2.5)$$

where λ_{min} is the smallest eigenvalue.

The difference between D_c and D_{max} is that, D_c is an absolute measure of the degree of multicollinearity. It is a fixed value for a data set, which does not change with any manipulation or technique used to analyze the data. It depends solely on the structure of the data set in the factor space. On the other hand, D_{max} is a relative measure in the sense that it depends on the orientation of the axes of the variable vectors. Its value changes with the data handling technique such as ridge regression. Simply speaking D_c is an absolute measure and D_{max} is a relative measure of the degree of multicollinearity whose value is reduced by ridge regression. It is always true that $D_c \geq D_{max}$, where in the special case, when one of the variable vectors lies on the major principal axis of the data set, the equal sign holds.

The advantage of the V_{max} measure of multicollinearity is that it is a linear scale while the D-measure is not. However, the D-measure gives a range of 0 to 1 for no collinearity to perfect collinearity, and is the same type of measure as a correlation coefficient. These two measures have different meanings and both should be provided by the analyst in order to assess the condition of the data on hand.

When a data set has a high degree of multicollinearity it is characterized by the fact that the smallest eigenvalue of its correlation matrix (X'X) is much smaller than unity (Hoerl and Kennard, 1970a). This condition generates at least the six following harmful effects on the OLS estimates of β.

(i) The estimated regression coefficients will have very large sampling variance.

The total sampling variance for OLS estimates is given as follows (Hoerl and Kennard 1970a):

$$\Sigma Var(\hat{\beta}) = \sigma^2 Tr(X'X)^{-1} \quad (2.6)$$

$$= \sigma^2 \Sigma \frac{1}{\lambda_i} > \sigma^2/\lambda_{min} \quad (2.7)$$

Where, and hereafter, Σ represents the summation from $i = 1$ to $i = p$, the number of predictors, and λ_{min} is the smallest eigenvalue of the correlation matrix (X'X),[3] which approaches zero for data sets approaching perfect collinearity. Therefore, the total variance; and, hence, the sampling variance of the

[3] In ridge regression (X'X) represents the correlation matrix, see chapter III.

individual predictors can be very large and may approach infinity when the data approaches perfect collinearity.

From a geometrical point of view, the total variance can be regarded as the squared distance from the estimated coefficient vector, $\hat{\beta}$, to the true vector β, and large total variance means that the distance from $\hat{\beta}$ to β is too long; hence, the greater the collinearity the longer the $\hat{\beta}$ to β distance.

(ii) The estimated coefficient vector $\hat{\beta}$ is far too long in general for data sets with high degree of multicollinearity.

In general, the expected square length of the estimated coefficient vectors $\hat{\beta}$ are far too long in ill-conditioned (high D-value) data sets as Hoerl and Kennard (1970a) have pointed out. That is:

$$
\begin{aligned}
\Sigma(\hat{\beta}'\hat{\beta}) &= \beta'\beta + \sigma^2\Sigma\frac{1}{\lambda_i} \\
&> \beta'\beta + \sigma^2/\lambda_{min} \qquad (2.8)
\end{aligned}
$$

When the degree of multicollinearity increases λ_{min} decreases, and the situation deteriorates such that the estimated coefficient vector $\hat{\beta}$ is far too long.

It may be concluded from properties (i) and (ii) that for an ill-conditioned data set there is a high probability that the estimated coefficient vector, $\hat{\beta}$, will be "off" not only in magnitude but also in direction; and this in turn leads to the conclusion that the estimated coefficients will be overly sensitive to the sampling error or model misspecification due to the large variance.

Under these conditions the inflated β estimates of predictor equations will not accurately reflect the net relative importance, of the variables under consideration. By the same token some non-significant coefficients may in fact be interpreted as important ones because they will be spuriously high. It is not unknown, for example, in situations where multicollinearity is a problem for the estimated coefficient to be larger than the correlation coefficient between the predictor and the dependent variable. Indeed, OLS regression coefficients may sometimes be larger than one.

(iii) When an estimation data set is ill-conditioned OLS regression models do not provide accurate prediction even when the R^2 level is high.

This is a well known fact among data analysts. In fact, from the above discussion about the regression coefficients estimated from ill-conditioned data we know that such coefficients are unstable and unreliable; and therefore cannot produce an accurate prediction. However, this conclusion can be formulated explicitly as follows.

The forecasting error variance, σ_f^2, of a multiple linear regression model, $y = x\beta + \varepsilon$, can be written as:

$$\sigma_f^2 = \sigma^2 \; [1 + \underset{\sim}{x}'V\underset{\sim}{x}] \qquad (2.9)$$

where $\underset{\sim}{x}$ is a (p x 1) column vector of a single observation of the predictor variables, X's; V is the variance-covariance matrix of the estimated regression coefficients, which is $(X'X)^{-1}$ for OLS estimates, and σ^2 is the variance of the random error in y. Therefore, for OLS estimated forecasting error variance can be expressed as:

$$\hat{\sigma}_f^2 = \hat{\sigma}^2 \; [1 + x'(X'X)^{-1}x]$$

$$= \hat{\sigma}^2 \; [1 + x^{*\prime} \Lambda^{-1} x^*]$$

$$= \hat{\sigma}^2 \; [1 + \Sigma \frac{x_i^{*2}}{\lambda_i}] \qquad (2.10)$$

or
$$\hat{\sigma}_f^2 = \frac{y'y(1-R^2)}{n-p} \; [1 + \Sigma \frac{x_i^{*2}}{\lambda_i}] > \frac{y'y(1-R^2)}{n-p} \frac{x_p^{*2}}{\lambda_{min}} \qquad (2.11)$$

where $x^* = P'x_o$ and P is the orthogonal transformation matrix of $(X'X)$, such that $P'(X'X)P = \Lambda$, the eigenvalue matrix of $(X'X)$, and $P'P = PP' = I$. From the above expression of forecasting error variance, it is clear that (a) the forecasting error variance can be very large for ill-conditioned data for which λ_{min} is very small, and (b) it is more sensitive to the smallest eigenvalue, λ_{min}, than to R^2. This is why in OLS regression, when the problem has a high degree of multicollinearity, high R^2 does not produce accurate prediction.

(iv) The OLS estimates, $\hat{\beta}$'s, are sensitive to sampling fluctuation when the data set is ill-conditioned.

We have seen clearly (property 1) that for an ill-conditioned data set the OLS estimates have very large sampling variance due to their inverse relationship with the eigenvalues (see Eq. 2.7). The sampling variance will approach infinity when the data set approaches perfect collinearity. Essentially, this implies that the OLS estimates are sensitive to sampling fluctuation.

(v) The estimated regression coefficients might be highly dependent and therefore are error prone.

As has been pointed out earlier, when the data set has a high degree of multicollinearity, the estimates have large variance and covariance. Experience has shown that this generally produces some highly correlated estimates. This is especially true for those predictors that are highly correlated. Due to this high dependency, if one is erroneous, the other will also be erroneous. It is generally true that when two predictors are highly positively correlated, the estimated coefficients are highly negatively correlated; that is, if one coefficient is over-estimated, the other will be under-estimated. This high correlation between estimates, plus the large variance of the estimates clearly accounts for why the OLS estimates generally have too large a sum of squares $\hat{\beta}'\hat{\beta}$, and why they might even have the wrong sign.

(vi) The OLS estimates are sensitive to model misspecification when the data set is ill-conditioned.

It is well known that for orthogonal data sets, the regression coefficients $\hat{\beta}$'s are not affected by model misspecification due to the inclusion or exclusion of certain relevant variables. That is, the $\hat{\beta}$'s are invariant of the model specification. However, when the data set is ill-conditioned, the $\hat{\beta}$'s are highly correlated and consequently no longer invariant of the model specification. Thus, the regression coefficients of included variables might change drastically as some relevant variables are dropped from, or added to the model. To what extent these coefficients change is largely dependent on the importance of the variable, and the extent to which it affects the variance inflation factor of each variable in the predictor variable set.

From the above discussion, it is obvious that multicollinearity is a very undesirable problem in multiple regression and should be avoided whenever possible. However, in some fields of research, such as the social sciences and even in the natural sciences, it is impossible to impose control over the variables; multicollinearity persists due to environmental or physical constraints (Marquardt and Snee, 1975). In fact, empirically, orthogonal explanatory variables are virtually nonexistent. Since multicollinearity is a matter of degree, OLS procedures always produce estimates that are inflated, overly sensitive, unreliable and hard to interpret. The extent of the seriousness depends on the level of the multicollinearity of the problem. Thus, with OLS regression the key question is: At what level of multicollinearity are the results unacceptable? The subjective answers depend on the predilections of the researcher - a much too esoteric a consideration.

In the next chapter, we will see that the harmful effects produced by OLS procedure will be damped out in ridge regression through the introduction of a small parameter in the model.

CHAPTER III

THEORY (II)

A DESCRIPTION OF RIDGE REGRESSION

The Simple Ridge Regression

As has been pointed out earlier, ordinary least squares regression based on the minimum residual sum of squares criterion does not produce satisfactory results, because their acceptability depends on the degree of multicollinearity of the data set which worsens as the degree of multicollinearity increases. Hoerl and Kennard (1970a) have developed a promising technique, called "ridge regression", which is based on a minimum total mean square error criterion.

In ridge regression, the statistical model and its assumptions are the same as those for OLS regression as presented in the previous chapter. However, for convenience in development, the variables (the X's and y) are standardized so that X'X gives the correlation matrix of the predictors and X'y gives the vector of correlation coefficients of the dependent and independent variables.

In Hoerl and Kennard's ridge regression the criterion for measuring the goodness of an estimator $\hat{\beta}^*$ is the total mean square error (MSE)[4] function defined by

$$MSE = E[(\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta)] \qquad (3.1)$$

[4] There are two MSE criteria used in ridge regression, weighted and unweighted (see Hocking et. al., 1976). In this study only the unweighted one is used.

MSE is the sum of the mean square error of the individual regression coefficients, and it can be proved (Pindyck and Rubinfeld, 1976, p. 22, or see property vii in this chapter) that it can be decomposed into the sum of the total variance and the total squared bias of each regression coefficient, that is

$$MSE = \Sigma Var(\hat{\beta}^*_i) + Bias^2(\hat{\beta}^*) \tag{3.2}$$

From a geometrical point of view, MSE represents the expected squared distance between $\hat{\beta}^*$ and the true vector β, thus, a "good" estimator would be the one that minimizes this distance - that is, minimizes the MSE. Furthermore, since the ordinary least square estimator, $\hat{\beta}$, gives the minimum residual sum of squares; $\hat{\beta}^*$, the estimator based on the minimum MSE criterion, will give an inflated residual sum of squares, which can be written as

$$\begin{aligned} \phi = \varepsilon'\varepsilon &= (y - X\hat{\beta}^*)' \ (y - X\hat{\beta}^*) \\ &= (y - X\hat{\beta})' \ (y - X\hat{\beta}) + (\hat{\beta}^* - \hat{\beta})' \ X'X(\hat{\beta}^* - \hat{\beta}) \\ &= \phi_{min} + \phi_o(\hat{\beta}^*) \end{aligned} \tag{3.3}$$

Where

$$\phi_o(\hat{\beta}^*) = (\hat{\beta}^* - \hat{\beta})' X'X(\hat{\beta}^* - \hat{\beta}) \tag{3.4}$$

is the inflation in the residual sum of squares. Geometrically it gives the surface of a family of hyperellipsoids centered at $\hat{\beta}$, which is the OLS estimate of β.

The so called ridge estimator, $\hat{\beta}^*$, is the one which minimizes the squared length of the coefficient vector subjected to constant inflation in the residual sum of squares, $\phi_o(\hat{\beta}^*)$, which is determined by the minimum MSE. That is, $\hat{\beta}^*$

is a solution to minimizing $\hat{\beta}^{*\prime}\hat{\beta}^*$ subjected to

$$(\hat{\beta}^* - \hat{\beta})' X'X (\hat{\beta}^* - \hat{\beta}) = \phi_o$$ (a constant determined by the minimum MSE).[5]

The solution of this problem is obtained by minimizing the Lagrangian function

$$f(\hat{\beta}^*) = \hat{\beta}^{*\prime}\hat{\beta}^* + \tfrac{1}{k}[(\hat{\beta}^* - \hat{\beta})'X'X(\hat{\beta}^* - \hat{\beta}) - \phi_o] \quad (3.5)$$

where 1/k is the Lagrangian multiplier. At minimum $f(\hat{\beta}^*)$, we have

$$\frac{\partial}{\partial\hat{\beta}^*} f(\hat{\beta}^*) = 2\hat{\beta}^* + \tfrac{1}{k} 2X'X(\hat{\beta}^* - \hat{\beta}) = o$$

$$(X'X + kI)\hat{\beta}^* = X'X\hat{\beta} = X'y$$

therefore

$$\hat{\beta}^* = (X'X + kI)^{-1} X'y \quad (3.6)$$

where k is a parameter determined by Eq (3.4), which is in turn determined by minimum MSE[6] (see Hoerl and Kennard, 1970a).

Graphically, the geometrical relation of the ridge procedure in two dimensional parameter space can be depicted

---

[5] In the Variance Normalization Simple Ridge Regression proposed by this study, ϕ_o is determined indirectly by normalizing the average VIF.

[6] Due to the cyclic relationship between k, ϕ_o and $\hat{\beta}^*$, minimum MSE alone cannot complete the ridge procedure without using a different method to pre-determine k or ϕ_o. Therefore, in practice, numerous methods have teen devised for determining k. However, none of these can guarantee to give the true k which minimizes MSE.

by Figure (1). Where the ellipse is the hyperellipsoid of

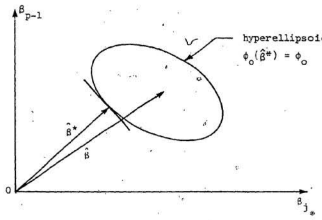


Figure 1: The Geometrical Relationship of the OLS Estimate, the Ridge Estimate, and the Inflation in the Residual Sum of Squares.

$\phi_o(\hat{\beta}^*) = \phi_o$ = constant, and the ridge estimator $\hat{\beta}^*$ is a vector which is shortest among those that have constant inflated residual sum of squares ϕ_o.

From the above argument we have seen that the minimum square length of the coefficient vector, or the minimum sum of squared regression coefficients determine the form of the ridge estimator and the minimum MSE determines the value of the parameter k. It should be noted that the form of the ridge estimator, i.e.,

$$\hat{\beta}^* = (X'X + kI)^{-1} X'y$$

is the same as that derived by Lindly and Smith (1972) using

Bayesian methods under the assumption of an exchangeable prior distribution for β.

The Properties of Simple Ridge Regression

Most of the properties of ridge regression have been thoroughly discussed by many researchers such as Hoerl and Kennard (1970a), and Marquardt (1970). Here, some sixteen properties of ridge regression are summarized. In several places the observations of others have been extended.

(i) Ridge regression gives a shorter regression coefficient vector than that of OLS regression.

Proof: For OLS regression, since $\hat{\beta} = (X'X)^{-1}X'y$, we have

$$\hat{\beta}'\hat{\beta} = y'X(X'X)^{-2}X'y = y'XP\Lambda^{-2}P'X'y$$

$$= y'X^*\Lambda^{-2}X^{*\prime}y = \Sigma \frac{(X^{*\prime}y)_i^2}{\lambda_i^2} \qquad (3.7)$$

Where $X^* = XP$ and P is the eigenvector matrix of $X'X$ which is an orthogonal transformation matrix satisfying $P'P = PP' = I$, and Λ is the eigenvector matrix of $X'X$ with eigenvalues $\lambda_1, \lambda_2, \ldots \ldots \lambda_p$, satisfying $X'X = P\Lambda P'$.

For ridge regression, $\hat{\beta}^* = (X'X + kI)^{-1}X'y$, and

$$\hat{\beta}^{*\prime}\hat{\beta}^* = y'X(X'X + kI)^{-2}X'y$$

$$= y'X^*(\Lambda + kI)^{-2}X^{*\prime}y$$

$$= \Sigma \frac{(X^{*\prime}y)_i^2}{(\lambda_i + k)^2} \qquad k \geq 0 \qquad (3.8)$$

From the above two relations, it is obvious that, $\hat{\beta}^{*\prime}\hat{\beta}^* \leq \hat{\beta}'\hat{\beta}$; hence, ridge regression $(k > 0)$ gives a shorter regression coefficient vector.

(ii) The ridge estimator $\hat{\beta}^*$ is a linear transformation of the OLS estimator $\hat{\beta}$, which is

$$\hat{\beta}^* = Z\hat{\beta} \tag{3.9}$$

Where $Z = (X'X + kI)^{-1}X'X$ or $Z = P\,\mathrm{Diag}\,[\lambda_i/(\lambda_i + k)]P'$, when it is expressed in terms of the eigenvalues of $X'X$, and where $\mathrm{Diag}\,[\lambda_i/(\lambda_i + k)]$ represents a diagonal matrix with ith diagonal element $\lambda_i/(\lambda_i + k)$.

Proof: The ridge estimator

$$\begin{aligned}\hat{\beta}^* &= (X'X + kI)^{-1}X'y \\ &= (X'X + kI)^{-1}X'(X\hat{\beta} + \hat{\varepsilon}) \\ &= (X'X + kI)^{-1}X'X\hat{\beta} \end{aligned} \tag{3.10}$$

where we have been using the fact that, in OLS regression, the residual vector is perpendicular to all independent variables X's, that is $X'\hat{\varepsilon} = 0$. Therefore, the linear transformation matrix is

$$Z = (X'X + kI)^{-1}X'X \tag{3.11}$$

and when expressed in terms of the eigenvalues of $X'X$, we have

$$\begin{aligned} Z &= P(\Lambda + kI)^{-1}\Lambda P' \\ &= P\,\mathrm{Diag}\,[\lambda_i/(\lambda_i + k)]P' \end{aligned} \tag{3.12}$$

where $\mathrm{Diag}\,[\lambda_i/(\lambda_i + k)] = \begin{pmatrix} \lambda_1/(\lambda_1 + k) & & & 0 \\ & \lambda_2/(\lambda_2 + k) & & \\ & & \ddots & \\ 0 & & & \lambda_p/(\lambda_p + k) \end{pmatrix}$

(iii) The ridge estimator $\hat{\beta}^*$ has a variance - covariance matrix

$$\mathrm{cov}(\hat{\beta}^*) = \sigma^2[\mathrm{VIF}] \tag{3.13}$$

where [VIF] is the variance inflation factor matrix, and

$$[VIF] = (X'X + kI)^{-1}X'X(X'X + kI)^{-1} \qquad (3.14)$$

or, when expressed in terms of the eigenvalues of the correlation matrix X'X

$$[VIF] = P \text{ Diag } [\lambda_i/(\lambda_i + k)^2]P' \qquad (3.15)$$

where Diag $[\lambda_i/(\lambda_i + k)^2]$ is a diagonal matrix with $\lambda_i/(\lambda_i + k)^2$ as its ith element.

Proof: Since the ridge estimator is a linear transformation of $\hat{\beta}$, i.e., $\hat{\beta}^* = Z\hat{\beta}$

we have

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

Therefore

$$[VIF] = (X'X + kI)^{-1}X'X(X'X + kI)^{-1}.$$

When it is expressed in terms of the eigenvalues of X'X, we apply $X'X = P\Lambda P'$, and obtain

$$\begin{aligned} [VIF] &= P(\Lambda + kI)^{-1}\Lambda(\Lambda + kI)^{-1}P' \\ &= P \text{ Diag } [\lambda_i/(\lambda_i + k)^2]P' \end{aligned}$$

(iv) Ridge regression produces smaller variances of the regression coefficients than that of OLS regression; however it does not necessarily reduce the covariance or correlation between them.

Proof: The variance - covariance matrix of ridge estimates is

$$\text{cov}(\hat{\beta}^*) = \sigma^2[VIF]$$

where

$$[VIF] = P \text{ Diag } [\lambda_i/(\lambda_i + k)^2]P'$$

is the variance - covariance inflation factor matrix with elements

$$VIF_{jl} = \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} p_{ji}p_{li} \quad \text{for all } j,l = 1,2,\ldots p$$

and where p_{ji}, which can be positive or negative, represents the ith element of the jth eigenvector in orthogonal transformation matrix P. Since $k > 0$ for ridge regression and $k = 0$ for OLS regression, the variances

$$VIF_{jj} = \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} p_{ji}^2 \quad \text{for all } j = 1, 2, \ldots p \quad (3.16)$$

are reduced by the ridge procedure. However, the covariances

$$VIF_{jl} = \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} p_{ji}p_{li} \quad \text{for } j \neq l \text{ and all } j,l = 1,2,\ldots p \quad (3.17)$$

can be inflated or deflated due to the uncertainty in the sign of "$p_{ji}p_{li}$", as compared to that of OLS regression.

From the above, it is clear then, that the correlation between $\hat{\beta}_j^*$ and $\hat{\beta}_l^*$, which is

$$r_{jl}^* = \frac{VIF_{jl}}{\sqrt{VIF_{jj}}\sqrt{VIF_{ll}}}$$

can also be inflated or deflated.

Comments: 1. When a simple ridge procedure is used, it's effects on the correlation between the regression coefficients should be checked and the analyst should interpret the estimated coefficients with caution. If the correlation coefficients are significantly different from zero, the statistical control of the variables will not have been achieved.

2. In a generalized ridge regression therefore, one should look not only at the variance but also at the covariance in order to reduce the correlation coefficients between the regression coefficients. This generalized ridge procedure called the "Generalized Normalization Ridge Regression" is under development.

(v) The total variance of ridge estimates

$$\Sigma \mathrm{Var}(\hat{\beta}_i^*) = \sigma^2 \Sigma \lambda_i/(\lambda_i + k)^2 \tag{3.18}$$

is a monotonic decreasing function of k.

Proof: The variance - covariance matrix (property iii) of ridge estimates is

$$\mathrm{cov}(\hat{\beta}^*) = \sigma^2 P \,\mathrm{Diag}\, [\lambda_i/(\lambda_i + k)^2]P'$$

Therefore, the total variance of the estimates is

$$\begin{aligned} \Sigma \mathrm{Var}(\hat{\beta}_i^*) &= \mathrm{tr}\ \mathrm{cov}(\hat{\beta}^*) \\ &= \sigma^2 \mathrm{tr}\ P\, \mathrm{Diag}\ [\lambda_i/(\lambda_i + k)^2]P' \\ &= \sigma^2 \mathrm{tr}\ \mathrm{Diag}\ [\lambda_i/(\lambda_i + k)^2]P'P \\ &= \sigma^2 \Sigma\ \lambda_i/(\lambda_i + k)^2 \end{aligned}$$

It is obvious then, since k and all λ_i for i = 1,2,....p, are positive values, the total variance is a monotonic decreasing function of k.

Comments: 1. The total variance has a range of $(\sigma^2 \Sigma \frac{1}{\lambda_i}, 0)$ when k varies from 0 to infinity. This means that the total variance reduces to zero when k approaches infinity. However, at large k, the variance reduces at a much slower rate.

2. The decreasing rate at $k \to 0$, that is

$$\lim_{k \to o} \left| \frac{d}{dk} \Sigma \mathrm{Var}(\hat{\beta}_i^*) \right| = 2\sigma^2 \Sigma \frac{1}{\lambda_i^2} \geqslant \frac{2\sigma^2}{\lambda_{min}} \tag{3.19}$$

can be very large for problems with a high degree of multicollinearity. This implies that ridge regression is more effective for problems with a high degree of multicollinearity.

(vi) The ridge estimator $\hat{\beta}^*$ is a negatively biased estimator with a bias given by

$$\mathrm{Bias}(\hat{\beta}^*) = -k(X'X + kI)^{-1}\beta \qquad (3.20)$$

or

$$\mathrm{Bias}(\hat{\beta}^*) = -k\,P\,\mathrm{Diag}\,[1/(\lambda_i + k)]P'\beta \qquad (3.21)$$

when it is expressed in terms of the eigenvalues of the correlation matrix.

Proof:

$$\begin{aligned}\mathrm{Bias}(\hat{\beta}^*) &= E(\hat{\beta}^*) - \beta = Z\beta - \beta \\ &= [(X'X + kI)^{-1}X'X - I]\beta \\ &= -k(X'X + kI)^{-1}\beta\end{aligned}$$

In terms of the eigenvalues, by substituting, $X'X = P\Lambda P'$ and $P'P = I$, we have

$$\mathrm{Bias}(\hat{\beta}^*) = -k\,P\,\mathrm{Diag}[1/(\lambda_i + k)]P'\beta$$

Comment: 1. It is obvious that the bias produced by ridge regression is a function of an unknown population regression coefficient vector β, and hence the bias cannot be calculated.

2. Since $\hat{\beta}^*$ is negatively biased, if we are able to prove $\hat{\beta}^*$ or $E(\hat{\beta}^*)$ is significant, then the true value β, must be significant.

(vii) The total square bias of ridge estimates

$$\mathrm{Bias}^2(\hat{\beta}^*) = k^2\Sigma\frac{\alpha_i^2}{(\lambda_i + k)^2} \qquad (3.22)$$

is a monotonic increasing function of k with a range of $(0, \Sigma\beta_i^2)$ for k = 0 to infinity.

Proof: From property (vi) we have the square bias of the estimates

$$\text{Bias}^2(\hat{\beta}^*) = k^2\beta'P\,\text{Diag}[1/(\lambda_i + k)]\,\text{Diag}[1/(\lambda_i + k)]P'\beta$$

$$= k^2\alpha'\text{Diag}[1/(\lambda_i + k)^2]\alpha$$

$$= k^2\,\Sigma\frac{\alpha_i^2}{(\lambda_i + k)^2}.$$

where $\alpha = P'\beta$ is the true regression coefficient vector in canonical form $y = X^*\alpha + \varepsilon$, and where $X^* = XP$. It is obvious therefore, that $\text{Bias}^2(\hat{\beta}^*)$ is a monotonic increasing function of k. Further, since

$$\lim_{k\to o}\text{Bias}^2(\hat{\beta}^*) = \lim_{k\to o}k^2\Sigma\frac{\alpha_i^2}{(\lambda_i + k)^2} = 0$$

and

$$\lim_{k\to\infty}\text{Bias}^2(\hat{\beta}^*) = \lim_{k\to\infty}k^2\Sigma\frac{\alpha_i^2}{(\lambda_i + k)^2}$$

$$= \lim_{k\to\infty}\Sigma\frac{\alpha_i^2}{(\frac{\lambda_i}{k} + 1)^2}$$

$$= \Sigma\alpha_i^2$$

$$= \Sigma\beta_i^2$$

the total square bias of ridge estimates has a range of $(0, \Sigma\beta_i^2)$.

Comment: The total square bias depends on the unknown population regression coefficient vector β; therefore, it can never be obtained.

(viii) The ridge estimator $\hat{\beta}^*$ has a total mean square error of

$$\text{MSE} = \sigma^2 \, \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \, \Sigma \frac{\alpha_i^2}{(\lambda_i + k)^2} \qquad (3.23)$$

Proof: The total mean square error

$$\begin{aligned} \text{MSE} &= E(\hat{\beta}^* - \beta)^2 = E\left\{[\hat{\beta}^* - E(\hat{\beta}^*)] + [E(\hat{\beta}^*) - \beta]\right\}^2 \\ &= E[\hat{\beta}^* - E(\hat{\beta}^*)]^2 + [E(\hat{\beta}^*) - \beta]^2 \\ &= \text{tr cov}(\hat{\beta}^*) + \text{Bias}^2(\hat{\beta}^*) \\ &= \sigma^2 \text{tr } P \,\text{Diag}[\lambda_i/(\lambda_i + k)^2]P' + k^2 \, \alpha' \text{Diag}[1/(\lambda_i + k)^2]\, \alpha \\ &= \sigma^2 \, \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \, \Sigma \frac{\alpha_i^2}{(\lambda_i + k)^2} \end{aligned}$$

Comments: 1. From this relation, it is obvious that the total mean square error of ridge regression depends on the true unknown parameters σ^2 and α's. Therefore, the total mean square error cannot be obtained.

2. The first component of MSE is the total variance of estimation, i.e.

$$\Sigma \text{VAR}(\hat{\beta}_i^*) = \sigma^2 \, \Sigma \frac{\lambda_i}{(\lambda_i + k)^2}$$

which describes the random portion of the error, while the second component is the square bias, i.e.

$$\text{Bias}^2(\hat{\beta}^*) = k^2 \Sigma \frac{\alpha_i^2}{(\lambda_i + k)^2}$$

which is the systematic portion of the error.

(ix) Ridge regression gives minimum distance between $\hat{\beta}^*$ and the true vector β, which in this sense makes $\hat{\beta}^*$ a better estimator than that of $\hat{\beta}$, the OLS estimator.

Proof: The ridge regression criterion demands an estimation procedure minimizing the mean square error. From a geometrical point of view MSE is the squared distance from $\hat{\beta}^*$ to β; therefore, the ridge estimator gives a minimum distance from $\hat{\beta}^*$ to the true vector β.

Comment: Since the MSE of ridge regression depends on the unknown parameters σ^2 and α, the minimum MSE or the minimum distance from $\hat{\beta}^*$ to β cannot be obtained.

(x) Ridge regression inflates the residual sum of squares of

$$\phi_0 = k^2\hat{\beta}^{*\prime}(X'X)^{-1}\hat{\beta}^* \tag{3.24}$$

or

$$\phi_0 = k^2\Sigma\hat{\alpha}_i^{*2}/\lambda_i \tag{3.25}$$

where $\hat{\alpha}^* = P'\hat{\beta}^*$ is the ridge regression coefficient vector in a factor space defined by orthogonal transformation $X^* = XP$, in which the multiple linear regression model in canonical form is given by $y = X^*\alpha + \varepsilon$.

Proof: Since $\hat{\beta}^* = Z\hat{\beta}$ or $\hat{\beta} = Z^{-1}\hat{\beta}$ we have

$$\begin{aligned}\hat{\beta}^* - \hat{\beta} &= (I - Z^{-1})\hat{\beta}^* \\ &= [I - (X'X)^{-1}(X'X + kI)]\hat{\beta}^* \\ &= -k(X'X)^{-1}\hat{\beta}^*\end{aligned}$$

Therefore

$$\begin{aligned}\phi_0 &= (\hat{\beta}^* - \hat{\beta})'X'X(\hat{\beta}^* - \hat{\beta}) \qquad \text{Eq. (3.4)} \\ &= k^2\hat{\beta}^{*\prime}(X'X)^{-1}\hat{\beta}^*\end{aligned}$$

In terms of eigenvalues, by substituting $X'X = P\Lambda P'$ and $P'\hat{\beta}^* = \hat{\alpha}^*$ we have

$$\phi_0 = k^2\hat{\alpha}^{*\prime}\Lambda^{-1}\hat{\alpha}^* = k^2\Sigma\hat{\alpha}_i^{*2}/\lambda_i$$

(xi) Ridge regression produces a smaller multiple R square than that of OLS regression which can be expressed as

$$R^2 = \hat{\beta}^{*\prime}X'X\hat{\beta}^* \tag{3.26}$$

$$= \hat{\beta}^{*\prime}X'y - k\hat{\beta}^{*\prime}\hat{\beta}^* \tag{3.27}$$

or, when expressed in terms of the eigenvalues of X'X, as

$$R^2 = \Sigma\lambda_i\hat{\alpha}_i^{*2} \tag{3.28}$$

Proof: The multiple R square

$$R^2 = \frac{RSS}{TSS} = \frac{\hat{y}^{*\prime}\hat{y}^*}{y'y} = \frac{\hat{\beta}^{*\prime}X'X\hat{\beta}^*}{y'y} = \hat{\beta}^{*\prime}X'X\hat{\beta}^*$$

where we have been using $y'y = 1$, since in ridge regression, all variables are standardized to give unit length. Further, by substituting $\hat{\beta}^* = (X'X + kI)^{-1}X'y$ and using a little algebra, we have

$$R^2 = \hat{\beta}^{*\prime}X'X\hat{\beta}^* = \hat{\beta}^{*\prime}X'X(X'X + kI)^{-1}X'y$$

$$= \hat{\beta}^{*\prime}[(X'X + kI) - kI](X'X + kI)^{-1}X'y$$

$$= \hat{\beta}^{*\prime}X'y - k\hat{\beta}^{*\prime}\hat{\beta}^*$$

And by using $X'X = P\Lambda P'$ and $P'\hat{\beta}^* = \hat{\alpha}^*$ we obtain the expression in terms of the eigenvalues of X'X, i.e.

$$R^2 = \hat{\beta}^{*\prime}X'X\hat{\beta}^* = \hat{\alpha}^{*\prime}\Lambda\hat{\alpha}^* = \Sigma\lambda_i\hat{\alpha}_i^{*2}$$

From Eq. (3.27), it can be seen clearly that, for OLS regression, the multiple R^2 is greater than that of ridge regression, since for OLS regression we have $k = 0$ and $|\hat{\beta}| > |\hat{\beta}^*|$.

Comment: The R^2 for ridge regression is a monotonic decreasing function of k. This can be seen from the fact R^2 can be expressed as

$$R^2 = \Sigma\frac{\lambda_i}{(\lambda_i + k)^2}(X^{*\prime}y)_i^2 \tag{3.29}$$

which is obtained from substituting $\hat{\alpha}_i^* = (X^{*\prime}y)_i/(\lambda_i + k)$ into Eq. (3.28).

(xii) The ridge estimate, $\hat{\beta}^*$, is less sensitive to sampling fluctuation as compared to the OLS estimate.

From property (iv) we know that the ridge procedure produces smaller sampling variance. Essentially, this implies that the ridge estimate is less sensitive to sampling fluctuation.

(xiii) The ridge estimate produces a more accurate prediction equation than the OLS regression procedure, if the bias introduced is not too large.

Proof: It is well known that for unbiased estimates the variance of the forecasting error is

$$\sigma_f^2 = \sigma^2[1 + \underset{\sim}{x}'V\underset{\sim}{x}]$$

where $\sigma^2 V$ is the variance - covariance matrix of the estimated parameters. For a biased estimate, such as a ridge estimate, the square bias should be added to the forecasting error variance; that is

$$\begin{aligned}\sigma_f^{*2} &= \sigma_f^2 + \text{Bias}^2(\hat{\beta}^*)\\ &= \sigma^2[1 + \underset{\sim}{x}'V\underset{\sim}{x}] + \text{Bias}^2(\hat{\beta}^*)\end{aligned}$$

Therefore, when it is expressed explicitly, we have

$$\begin{aligned}\sigma_f^{*2} &= \sigma^2\left\{1 + \underset{\sim}{x}'P\ \text{Diag}[\lambda_i/(\lambda_i + k)^2]P'\underset{\sim}{x}\right\} + \text{Bias}^2(\hat{\beta}^*)\\ &= \sigma^2\left\{1 + \underset{\sim}{x}^{*\prime}\text{Diag}[\lambda_i/(\lambda_i + k)^2]\underset{\sim}{x}^*\right\} + \text{Bias}^2(\hat{\beta}^*)\\ &= \sigma^2\left[1 + \Sigma\frac{\lambda_i\ x_i^{*2}}{(\lambda_i + k)^2}\right] + \text{Bias}^2(\hat{\beta}^*) \qquad (3.33)\end{aligned}$$

For OLS estimates, we set k = 0 and derive the forecasting error variance as

$$\sigma_f^2 = \sigma^2\left[1 + \Sigma \frac{x_i^{*2}}{\lambda_i}\right] \tag{3.34}$$

which is much larger than σ_f^{*2} if the bias produced by ridge procedure is not large relative to the reduction in variance.

Comments: 1. The forecasting error variance σ_f^{*2} consists of two parts, the first part is the variance term, which describes the random portion of the error, and the second part (the bias term) describes the systematic portion of the error.

2. In the case where the bias term is relatively large, more accurate prediction can still be obtained by dividing the sample (if large enough) into two sets, one is used to estimate the biased parameter, and the other is used to estimate the bias in the prediction of the dependent variable. The accuracy of this empirically estimated bias in prediction can be improved by repeating the procedure with different ways of dividing the data set.

(xiv) There exists a wide range of, k, $0 < k < k_{max}$, which will give a set of ridge estimates $\hat{\beta}^*$; and which will produce smaller MSE than that of OLS estimates.

Proof: Define the effectiveness index (Eft) of ridge regression as the ratio of reduction in total variance to the total square bias introduced by the ridge regression; that is

$$\text{Eft} = \frac{\text{Reduction in total variance}}{\text{Bias}^2(\hat{\beta}^*)} \tag{3.35}$$

$$= \frac{\sigma^2 tr(X'X)^{-1} - \sigma^2 tr[VIF]}{\text{Bias}^2(\hat{\beta}^*)}$$

$$= \frac{\sigma^2 \left[\Sigma \frac{1}{\lambda_i} - \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} \right]}{k^2 \Sigma \frac{\alpha_i^2}{(\lambda_i + k)^2}} \tag{3.36}$$

Since the total variance of ridge regression is a monotonic decreasing function and $Bias^2(\hat{\beta}^*)$ is a monotonic increasing function of k, then Eft is a decreasing function of k. It can be proved easily that, the effectiveness of ridge regression has a range of infinity to zero when k varies from zero to infinity. Further, we have

$$\begin{aligned} MSE(\hat{\beta}) - MSE(\hat{\beta}^*) &= \Sigma Var(\hat{\beta}) - \Sigma Var(\hat{\beta}^*) - Bias^2(\hat{\beta}^*) \\ &= Eft \times Bias^2(\hat{\beta}) - Bias^2(\hat{\beta}^*) \\ &= (Eft - 1)\ Bias^2(\hat{\beta}^*) \end{aligned}$$

then, for any k which gives $Eft > 1$, we have

$$MSE(\hat{\beta}) - MSE(\hat{\beta}^*) > 0$$

That is, the ridge regression procedure produces smaller MSE. If we set $k = k_{max}$ (maximum k) for $Eft(k) = 1$, then all k's that are less than k_{max} would give smaller MSE.

Comments: 1. Any valid ridge procedure should produce an optimal k which is less than k_{max}.

2. The maximum k defined here is a function of unknown parameters σ^2 and α's, and hence the true value will never be obtained. However, if we use the OLS estimate of σ^2 and α's, we would obtain a conservative estimate of k_{max} (called $\hat{k}_{max}$), due to the fact that α's are generally over-estimated by OLS regression.

3. Since $\hat{k}_{max}$ is a conservative estimate, $k < \hat{k}_{max}$ is a sufficient but not necessary condition for a valid ridge procedure.

4. $\hat{k}_{max}$ and the maximum k (called k^v_{max}) defined by Vinod (1976, 1978) refer to the same theoretical maximum k, however empirically $\hat{k}_{max} > k^v_{max}$. That is $\hat{k}_{max}$ is a more accurate estimate of k^v_{max}.

5. The OLS estimate of the effectiveness index

$$\hat{Eft}(k) = \frac{\hat{\sigma}^2 \left[\Sigma \frac{1}{\lambda_i} - \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} \right]}{k^2 \, \Sigma \frac{\hat{\alpha}_i^2}{(\lambda_i + k)^2}}$$

may be used to indicate the performance of a ridge procedure. If $\hat{Eft}(k) > 1$, the ridge procedure is a valid one. However, as in comment (2), $\hat{Eft}(k)$ is a conservative estimate of the true effectiveness index, and therefore $\hat{Eft}(k) > 1$ is a sufficient not a necessary condition.

(xy) For any problem there exists a positive "optimal k" (called k_o) which gives a minimum MSE.

Proof*: Since

$$MSE = \Sigma Var(\hat{\beta}_i^*) + Bias^2(\hat{\beta}^*)$$

$$\frac{d}{dk} MSE = \frac{d}{dk} \Sigma Var(\hat{\beta}_i^*) + \frac{d}{dk} Bias^2(\hat{\beta}^*)$$

$$= -2\sigma^2 \, \Sigma \frac{\lambda_i}{(\lambda_i + k)^3} + 2k \, \Sigma \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3}$$

*This proof follows Kasarda and Shih (1977).

$$\doteq -2\ \Sigma \frac{\lambda_i}{(\lambda_i + k)^3} (\sigma^2 - k\alpha_i^2) \qquad (3.39)$$

Based on Rolle's Theorem (Widder 1963), since

$$\lim_{k\to o} \frac{d}{dk} MSE = -2\sigma^2\ \Sigma \frac{1}{\lambda_i^2} < 0$$

the optimal k is a positive value. Further since

$$\frac{d^2}{dk^2} MSE = 6\sigma^2\ \Sigma \frac{\lambda_i}{(\lambda_i + k)^4} + 2\ \Sigma \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^4} (\lambda_i - 2k) \quad (3.40)$$

and

$$\lim_{k\to o} \frac{d^2}{dk^2} MSE = 6\sigma^2\ \Sigma \frac{1}{\lambda_i^3} + 2\ \Sigma \frac{\alpha_i^2}{\lambda_i^2} > 0 \qquad (3.41)$$

According to the theorem of minimum (Widder 1963), this positive value leads to the conclusion that a minimum exist for MSE (Kasarda and Shih, 1977).

Comment: This is what Hoerl and Kennard (1970a) called the "existence theorem". It is true even for an orthogonal system for which the degree of multicollinearity, D, is zero. In this case, $\lambda_i \doteq 1$ for all i = 1,2,.....p. and we have

$$\frac{d}{dk} MSE = -2\ \Sigma \frac{\lambda_i}{(\lambda_i + k)^3} [\sigma^2 - k\alpha_i^2] = 0$$

$$\Sigma(\sigma^2 - k\alpha_i^2) = 0$$

$$p\sigma^2 - k\Sigma\alpha_i^2 = 0$$

$$\therefore\quad k = k_o = \frac{p\sigma^2}{\Sigma\alpha_i^2} = \frac{p\sigma^2}{\Sigma\beta_i^2}$$

This strange phenomenon has one important advantage and one important disadvantage. The advantage is that if we accept the minimum MSE criterion as a measure of the goodness of an estimator, for any problem, the OLS procedure is always inferior. The disadvantage is that it renders Hoerl and Kennard's ridge procedure based on minimum MSE rather absurd.

(xvi) For any problem the optimal k depends on the true regression coefficient vector, β, and the variance of the residual of the linear model, i.e. σ^2.

Proof: From Eq. (3.39), at minimum MSE

$$\frac{d}{dk}\,\mathrm{MSE} = -2\,\Sigma\,\frac{\lambda_i}{(\lambda_i + k)^3}\,(\sigma^2 - k\alpha_i^2) = 0 \qquad (3.43)$$

Although the explicit form of the optimal k cannot be solved from this complex non-linear equation, it is obvious that the optimal k is a function of the true regression coefficient vector, α or β, and the variance of the residual of the linear regression model, i.e. σ^2.

Comment: The multicollinearity problem arises from the interdependency among the explanatory variables not from the dependency between the dependent variable y, and the explanatory variables X's. Therefore, if the task of ridge regression is to reduce the harmful effect of multicollinearity, the optimal k should not depend on any parameter which depends on the y variable such as β or σ^2; that is, the optimal k should be a nonstochastic parameter.

The Optimal k

As has been defined in a previous section, the optimal k is the one which gives the minimum MSE for the data on hand,

and it has been pointed out that for any problem there is one optimal k, and a wide range of k, $0 < k < k_{max}$, which give smaller MSE than that of OLS regression. Unfortunately, the optimal k depends on the true regression coefficient vector, β, and the variance of the residual, σ^2, in the linear regression model. These two parameters are population parameters not universal constants, and due to this nature of the optimal k, it is impossible for it to be calculated. Instead, it has to be estimated from the sample data. So far more than fifteen methods have been described; for example, see Hoerl and Kennard (1970a), Hoerl, Kennard and Baldwin (1975), Vinod (1976), Obenchain (1975), Hocking et. al. (1976), McDonald and Galarneaus (1975), Kasarda and Shih (1977), Hemmerle (1975), Hemmerle and Brantle (1978), Guilkey and Murphy (1975), Lawless and Wang (1976), Allen (1974). Each of these methods has its own advantages and disadvantages. However, none can guarantee to give a better k or even a smaller MSE compared to that of OLS regression. This difficulty has unfortunately marred the superiority of ridge regression over OLS regression procedure. In the following section three distinctive methods of estimating the optimal k will be discussed.

1. Hoerl and Kennard's Ridge Trace Method. In Hoerl and Kennard's version of simple ridge regression, the optimal k is determined visually from the "ridge trace" which is the plot of $\hat{\beta}_i^*$'s and the residual sum of squares as functions of k. An example below is a ridge trace of a 10-factor

problem obtained from Hoerl and Kennard (1970b).

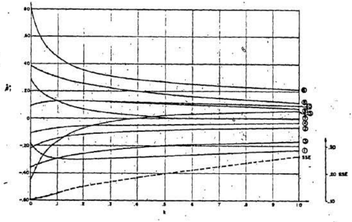


Figure 2: The Ridge Trace of the 10-Factor Problem from Hoerl and Kennard (1970b).

The ridge trace depicts the effect of multicollinearity on each of the regression coefficients. From the trace it can be seen that when k increases the effect of multicollinearity is dampened and the regression coefficients are stabilized. The optimal k is then selected visually at the region which starts to give stabilized regression coefficients.

In the Hoerl and Kennard (1970a) article four guidelines were suggested for the selection of the optimal k.

(1) At a certain value of k the system will stabilize and have the general characteristics of an orthogonal system.

(2) Coefficients will not have unreasonable absolute values with respect to the factors for which they represent rates of change.

(3) Coefficients with apparently incorrect signs at $k = 0$ will have been changed to the proper sign.

(4) The residual sum of squares will not have been inflated to an unreasonable value. It will not be large relative to the minimum residual sum of squares or large relative to what would be a reasonable variance of the process generating the data.

It is obvious that these four guidelines are vague, subjective, need prior knowledge of the regression coefficients and would prove very difficult to apply. Furthermore, the trace appears to be more stable at higher k and hence has a tendency to lead one to select a k that might be too high. Due to these drawbacks, the obtained "optimal k" cannot guarantee to give estimates that are better than the OLS ones. In spite of these limitations, the ridge trace is still a useful plot. It distinctively displays the characteristics of the explanatory data set, the effectiveness of ridge procedure in stablizing the regression coefficients, and can, also be used to check the optimal k estimated by using various methods to see if they fall in the stable region of the ridge trace as desired by a good estimator.

2. Kasarda and Shih's Method. As has been pointed out, theoretically the true optimal k is one which minimizes

the total mean square error of the estimates of regression coefficients. However, the MSE depends on the unknown true regression coefficients and the variance of the residuals (see property viii). Mathematically or technically it is not difficult to obtain the optimal k from Eq. (3.43), even for very high p, if the two parameters were known. Kasarda and Shih (1977) have argued that since the OLS estimates of σ^2 and β, under the normality assumption, are unbiased and consistent; and, further, since $\hat{\beta}$ has the minimum variance among all unbiased estimators, then the two OLS estimates, $\hat{\sigma}^2$ and $\hat{\beta}$, may be used to replace their true values in order to obtain the optimal k by minimizing the OLS estimate of MSE, which is written as

$$\hat{MSE}(\hat{\beta}^*) = \Sigma \hat{Var}(\hat{\beta}_i^*) + \hat{Bias}^2(\hat{\beta}^*)$$

$$= \hat{\sigma}^2 \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \Sigma \frac{\hat{\alpha}_i^2}{(\lambda_i + k)^2}$$

The validity of this method obviously rests on the validity of replacing σ^2 and α by their OLS estimates. The replacement of σ^2 is not problematic (Johnston 1972, p. 163), however the replacement of α (or β) by $\hat{\alpha}$ (or $\hat{\beta}$) definitely is, because the OLS estimates of α (or β) could be far off due to the high degree of multicollinearity in the problem (see harmful effect (ii) in chapter II). The replacement of α by $\hat{\alpha}$ would definitely produce too large a square bias in the estimate of MSE, and this in turn would produce an estimate of k (called k_s) smaller than the true optimal k (k_o), due to the fact that the variance component in MSE is a monotonic decreasing function and the square bias component is a monotonic increasing function with $\beta'\beta$ as its upper limit. Without a rigorous proof, this argument can be depicted graphically by Figure 3.

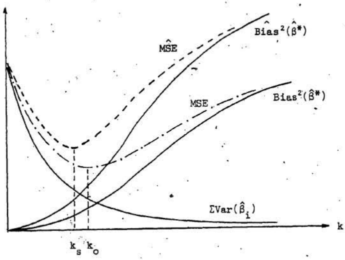


Figure 3: The mean square error function MSE and its OLS estimate, $\widehat{MSE}$, where k_s is Kasarda & Shih's k, and k_o is the true optimal k.

Summarized briefly, Kasarda and Shih's OLS estimation method have the following disadvantages:

(i) it produces a k which is an underestimate of the true optimal k, the higher the degree of multicollinearity, the worse the estimation; and

(ii) it produces a stochastic k, while optimal k should be nonstochastic due to the fact that multicollinearity is a nonstochastic problem caused by the interdependency in the predictors.

However, the first disadvantage, can also be regarded as an advantage from a different point of view; that is, it will never be an overestimate of k and thus produce too high a bias as some of the proposed method of estimating k do. In other words, Kasarda and Shih's method always produces an estimate with smaller MSE than that of OLS regression.

3. Vinod's Index of Stability Method. As pointed out earlier, the ridge estimator $\hat{\beta}^* = (X'X + kI)^{-1}X'y$ has a variance - covariance matrix of

$$\operatorname{cov}(\hat{\beta}^*) = \sigma^2 [\mathrm{VIF}]$$

where

$$[\mathrm{VIF}] = P \text{ Diag} \left[\frac{\lambda_i}{(\lambda_i + k)^2}\right] P'$$

is the variance inflation factor matrix. For a completely orthogonal system, $\lambda_1 = \lambda_2 = \ldots\ldots = \lambda_p = 1$, it can be easily seen that the VIF matrix is equal to a constant matrix with elements

$$\frac{\lambda_i}{(\lambda_i + k)^2} \quad \left(\text{i.e. } \frac{1}{(1 + k)^2}\right)$$

and therefore the matrix

$$\frac{p}{\Sigma \dfrac{\lambda_i}{(\lambda_i + k)^2}} \; P \text{ Diag} \left[\frac{\lambda_i}{(\lambda_i + k)^2}\right] P'$$

is an identity matrix.

For a non-orthogonal system the above property of the VIF will not be satisfied, and the absolute values of the elements will be large. This suggests a numerical measure, which Vinod (1976) called the Index of Stability of Relative

Magnitude (ISRM) of $\hat{\beta}^*$, defined by

$$ISRM = \Sigma\ [p\lambda_i/(\lambda_i + k)^2 S - 1]^2$$

where $$S = \Sigma\ \lambda_i/(\lambda_i + k)^2$$

This numerical measure represents the quantification of Hoerl and Kennard's concept of stability which will be zero for a completely orthogonal system. Since in ridge regression it is desired to minimize the effects of the non-orthogonality of the system, the index of stability (for short) should be minimized; that is, for optimal k, ISRM has a minimum value.

Due to the complexity of the stability function, it is impossible to solve for the optimal k as an explicit function of λ_i, and therefore it has to be solved graphically by plotting the ISRM as a function of k or by using an iterative approach.

This method may seem to be one of the best compared to most of the methods that have been proposed. However, it has not been very satisfactory in this study. To summarize it has at least the following advantages and disadvantages.

Advantages: (1) It quantifies Hoerl and Kennard's concept of stable region and estimates k objectively.

(2) It gives a more definitive k than the ridge trace method.

(3) It gives a non-stochastic k.

Disadvantages: (1) It was not proved that the optimal k gives the minimum mean square error as required by the criterion of ridge regression.

(2) There is no guarantee that the optimal k obtained will not be larger than k_{max} (see property xiv), as required by any valid ridge procedure.

(3) In some cases (see Appendix A), the index of stability gives more than one minimum point, while theoretically there should be only one optimal k that gives the minimum MSE (property xv).

CHAPTER IV

THEORY (III)

THE VARIANCE NORMALIZATION CRITERION

Introduction

From what have been discussed about the properties of ridge regression in chapter III: we have seen four dilemmas of ridge regression based on the minimum MSE criterion. That is (i) the theoretical value of optimal k is stochastic while it should be a nonstochastic one if the task of ridge regression is to reduce the harmful effect of multicollinearity, (ii) the true optimal k depends on population parameters and renders the problem unsolvable, (iii) even for an orthogonal system there is an optimal k at which the MSE is minimum, and (iv) the bias of the ridge estimates can never be obtained; and, thus the performance of ridge regression cannot be accurately evaluated. If we study the properties of ridge regression carefully, especially property (viii), (xv) and (xvi), it is clear that all these dilemmas stem from one source, the minimum MSE, due to the fact that because the mean square error is a function of β and σ^2, it is a stochastic function and cannot be evaluated accurately.

If we observe the minimum MSE criterion closely we would see that we might accomplish three tasks by using it with ridge regression. The level of accomplishment depends entirely on the size of the estimated k. In the remainder of this chapter we propose to use a weaker criterion in the sense that it is limited to the accomplishment of a single task;

namely, to reduce the effects of multicollinearity. Through the use of this criterion, called the variance normalization criterion, the first three of the above dilemmas would be avoided.

Analysis of the Problem

The total variance of the estimated regression coefficients based on the OLS procedure can be expressed as:

$$\Sigma\mathrm{Var}(\hat{\beta}) = \sigma^2 \mathrm{tr}(X'X)^{-1}$$

$$= \frac{\varepsilon'\varepsilon}{n-p}\,\mathrm{tr}(X'X)^{-1} \qquad (4.1)$$

where ε is the random error of y in the regression model $y = X\beta + \varepsilon$, n is the sample size, p is the number of explanatory variables and $\mathrm{tr}(X'X)^{-1}$ is the sum of the diagonal elements of $(X'X)^{-1}$, the inverse of the correlation matrix. From this expression it is obvious that the total variance depends on three independent factors, namely:

(i) The random error: This is the purely random part of the dependent variable y, and it generally consists of two parts, the measurement error ε_m, and the stochastic error ε_s, which can be regarded as the influence of the incompleteness of the designed model and some unknown inherent irreproducible fluctuation (Wannacott 1969, p. 17). These two errors ε_m and ε_s are assumed to be uncorrelated and have normal distribution with zero mean. The sum of the squares of the random error $\varepsilon'\varepsilon$ can therefore be expressed as

$$\varepsilon'\varepsilon = \varepsilon_m'\varepsilon_m + \varepsilon_s'\varepsilon_s \tag{4.2}$$

The random errors can be reduced but not eliminated.

(ii) The sample size n, or to be more specific, the degree of freedom of the sum of square error (n-p). The total variance may be reduced by increasing the sample size.

(iii) The Degree of Multicollinearity: The effect of the degree of multicollinearity enters through $tr(X'X)^{-1}$, which is the trace of $(X'X)^{-1}$ and it is the sum of the variance inflation factor of each variable due to its interrelationship with the rest of the explanatory variables. The inflation of variance by multicollinearity may be "normalized" by ridge regression developed by Hoerl and Kennard (1970a) if the k is constrained according to the variance normalization criterion as given in this chapter.

In any statistical procedure, it is desirable to have variance as low as possible. From the above discussion, it is obvious that for multiple linear regression, we may reduce the total variance of the estimated coefficients by improving the measurement, the specification of the model, the sample size and most important by reducing the inflation of variance due to multicollinearity, because as has been pointed out in chapter III, the total variance may be inflated to infinity (in the case discussed by Marquardt and Snee (1975), the VIF was as high as 6563). Due to the seriousness of the effect of multicollinearity, the analyst should locate their sources (see Mason, Gunet, and Webster, 1975) and try to eliminate them if physically possible, otherwise we have to resort to ridge regression to reduce the harmful effects of the multicollinearity problem. After all, to eliminate the causes is far better than to treat the symptom.

The ridge regression developed by Hoerl and Kennard (1970a) was originally intended to do just one task namely to reduce the inflation of variance of the estimates due to multicollinearity. This task is definitely a nonstochastic one (see the comment under property (xvi) in chapter III). However, use of the minimum MSE criterion for determining k is effectively to use an omnibus procedure. This is because ridge regression in suppressing the VIF to less than unity is also suppressing the variance attributable to other causes of variance - measurement error, model incompleteness or system misspecification and small sample size. It is the omnibus or multifunctional nature of the procedure which has forced the ridge regression procedure to be a stochastic one. This is why the theoretical value of optimal k based on minimum MSE is stochastic (depends on β and σ^2), and why even for an orthogonal system there is an optimal k which generates minimum MSE (see the comment under property (xv) in chapter III).

The Variance Normalization Criterion

From the above argument, it is obvious, therefore, if we want to limit the simple ridge procedure to perform the single task of reducing the inflation of variance due to multicollinearity, the lowest permissible total variance is $p\sigma^2$, which is the value for the best condition; namely, that present in an orthogonal system. That is,

$$\Sigma \mathrm{Var}(\hat{\beta}^*) = p\sigma^2$$

$$\sigma^2 \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} = p\sigma^2$$

and therefore

$$\frac{1}{p} \Sigma \frac{\lambda_i}{(\lambda_i + k)^2} = 1 \qquad (4.3)$$

Put into words we can say that in order to perform the single task of reducing variance inflation due to multicollinearity, and only multicollinearity, one should normalize the average variance inflation factor (VIF) such that it is equal to one. The value of k (called k_N) which satisfies this condition is the k which satisfies Eq. (4.3).

By using this procedure, if the resultant variance is still too large for practical application of the regression model; and if it is desirable to further reduce the variance, it should be accomplished through improvement in measurement error, model specification and sample size; and <u>not</u> by further suppression of the variance inflation factor.

<u>The Underlying Assumption, Limitations and Advantages</u>

It was stated earlier in this chapter that the ridge regression with the minimum MSE criterion <u>might</u> accomplish three tasks, and that the level of accomplishment depends entirely on the size of the estimated k. To be more specific, this is when the estimated k is larger than k_N, or the k value which normalized the average VIF equal to one. In the development of the normalization criterion, we did not and do not find it necessary to assume that k_N is always less than the optimal k. However, like any other method of estimating k, the variance normalization criterion has an underlying assumption that k_N is less than k_{max}, which is the k value for which the reduction in the total variance is equal to the total square bias introduced by the ridge procedure. In term of effectiveness index, Eft, k_{max} is the k value which gives Eft = 1 (see property xiv in chapter III).

With simple ridge regression, due to the crudeness of the process, the underlying assumption will not always be true. A Monte Carlo simulation experiment is desirable in order to evaluate where it stands. A further refinement, and generalization of this criterion, in order to achieve still better results, is definitely necessary. In spite of these limitations the normalization criterion has the following advantages:

(1) The parameter k_N can be calculated accurately and it is nonstochastic as required.

(2) The average variance inflation factor will never be suppressed.

(3) It is more conservative than some proposed methods.

(4) It helps to narrow down or even locate the source of variance in a model.

To illustrate the last advantage, let us assume there is a model with a very large data set; and further assume the model is well-specified based on information from other sources; then, after the data set is analyzed by using ridge regression with the variance normalization criterion, if the variance is still too large, its source most likely is from measurement error. This feature might become a helpful method to evaluate the crudeness of measurement in educational research.

CHAPTER V

EDUCATIONAL APPLICATION

Introduction

The purpose of this chapter is to demonstrate the superiority of ridge regression over OLS regression as claimed in the theoretical portion (chapter 2-4). The problem used for this purpose is the "human capital" problem, based on a modification and replication of Jencks' model (Bulcock et.al., 1974) in a Swedish context, through use of the Malmö data set.

Due to the fact that multicollinearity is most severe in the last stage of a structural equation model, the discussion here centers largely on the last stage of the model, although the whole model is analyzed for the sake of completeness. As stated earlier, the purpose of this chapter is to provide empirical support for the theoretical arguments about the superiority of simple ridge regression. Because this purpose is largely pedagogical, not substantive, the simplest - not necessarily the best - model was chosen. Thus, the interaction terms called for by resource conversion theory (Coleman 1971, Bulcock et.al., 1975, Fägerlind 1975) - an extension of human capital theory - were not included in the model used here for illustrative purposes. Furthermore, simple ridge regression is still not the perfect technique. Although the "noise" due to multicollinearity was reduced, the "distortion" (bias) may not be optimal. Therefore, at this point in time, there will be no statistical inference or claim about any fact or "truth of nature". In subsequent research, when better models and more perfect techniques, such as the Generalized Ridge Regression (GRR) based on the variance normalization criterion currently under development are used, then statistical inference about the "truth" will be stated.

The Malmö Data

The world famous Malmö data set is a longitudinal data set first collected in 1938 from all 1,544 grade three pupils in private and public schools of the city of Malmö in southern Sweden. The data gathering which was conducted in six different follow-up phases is summarized in Table 1.

Table 1 About Here

The details of the Malmö data set can be found in many articles such as those of the researchers that collected the data as given in Table 1. The Malmö data set has been widely used by economists, sociologists and educators to study human capital problems, and most recently by de Wolff and Van Slijpe (1973), Hause (1972, 1975), and Fägerlind (1975).

The Career Achievement Model

The career achievement process was first studied by Blau and Duncan (1967), and extended by Jencks et. al. in 1972. The model was replicated and further modified by Bulcock et. al. (1974), Fägerlind (1975) in the Swedish context by using the

Table 1

Phases in the Collection of the Malmö Data Set 1938-1973

Date of Collection	Type of Data	Size of Sample	Source of Data	Mode of Collection	Principal Researchers
1938	Group intelligence test	1544 (835 boys, 709 girls) (100%)	All third grade children in Malmö public and private schools	Pencil and paper test	Hallgren (1939)
	Demographic data		Taxpayers register Population registers, School class registers, and Social welfare register	Public records	
1942	Types of school to which students transferred, and scholastic ratings	440	All children transferred to junior secondary or higher school	Teacher ratings	Hallgren (1943)
1948	Social data, school marks, IQ test at maturity	613	All male respondents enrolled for military service	Military records, pencil and paper test	Husén (1947, 1948, 1950) Husén and Henricson (1951)

Table 1 (continued)

Date of Collection	Type of Data	Size of Sample	Source of Data	Mode of Collection	Principal Researchers
1958-65	Criminality data, social assistance data, and education data	104	Central criminal register Central welfare registers Malmö schools and central bureau of statistics	Public records	Husén, Emanuelsson, Fägerlind & Liljefors (1969)
	Income data	1236 (80.1%)	County tax departments	Public records	
	Social background data	1116 (81.2%)	Questionnaires	Mail	
1971-73	Adult education Data on occupations and working conditions	1077 (72%)	Questionnaire	Mail	Emanuelsson, Fägerlind, & Hartman (1973)
	Social welfare and criminality data Income data			Public records	
1974	2nd generation data on: IQ at maturity Expectations data School marks Type of school program		Military records	Data collection underway	Fägerlind

Malmö data. The analysis in this study is based mainly on Bulcock's model with two modifications, namely: (1) the interaction terms between variables were not included for the sake of simplicity; and (2) the outcome variable at each stage was regressed on all independent variables at that stage without hypothesizing as to whether each variable was going to effect the outcome or not. Any analysis is designed to find out the truth about nature, and if a relation does not exist, the results of a reliable analytical technique should indicate it explicitly. If we falsely hypothesize that a certain variable does not affect the outcome, the exclusion of that variable would produce a seriously biased estimate (Johnston 1972, pp. 168). Based on the seriousness of the effect of excluding any important or relevant variable from a regression model, econometric theory has pointed out that, when data and degrees of freedom permit, it is better to err on the side of including variables in regression analysis rather than excluding them (Johnston 1972, pp. 169). Therefore, the path diagram used in this study can be depicted as Figure 4.

Figure 4 About Here

Where FATHED(X_1) is father's education; FAMINC(X_2) is the family income (FAMINC included both father's and mother's income plus income from all other sources); FATHOCC(X_3) is father's occupation (a composite variable heavily dependent on occupational classification); FAMSZ(X_4) is the family size composed

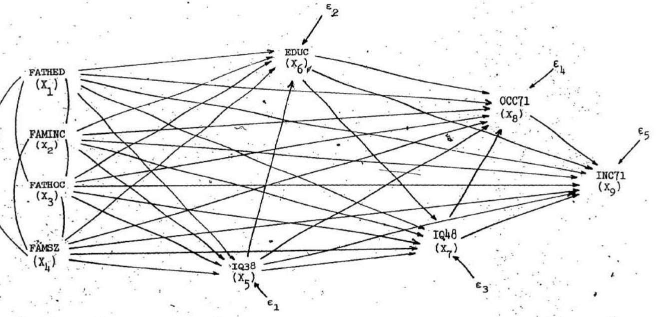


Figure 4: Path Diagram of the Malmö Model of the Socioeconomic Career where FATHED = father's education; FAMINC = family income; FATHOC = father's occupation; FAMSZ = family size; IQ38 = respondent's mental ability at age 10 (1938); EDUC = respondent's education level; IQ48 = respondent's mental ability at age 19 (1948); OCC71 = occupational status in 1971; INC71 = respondent's income in 1971.

solely of the number of siblings; IQ38(X_5) is the IQ score based on Hallgren's (1939) group intelligence test; EDUC(X_6) is the educational attainment measured on a four point scale; IQ48(X_7) is the mental ability at maturity (about age 19) based on the Swedish military intelligence test; OCC71(X_8) is the respondents' occupational status classified on a six point scale; and INC71(X_9) is the raw income data obtained from the central tax register and which included income from all sources. The details of these variables can be obtained from Fägerlind (1975).

The Analysis and the Results.

An OLS regression analysis and several simple ridge regression analyses were performed on the Malmö data set given in Bulcock et.al. (1974). The data is shown in Table 2.

Table 2 About Here

Although only four simple ridge regression methods were discussed in the theoretical chapters, seven methods were used in the analyses presented in this chapter. The three extra are: Hocking, Speed and Lynn's (1976) method, Lawless and Wang's (1976) method, and Hoerl, Kennard and Baldwin's (1975) method. In these methods, the estimators for the optimal k are:

Hocking et.al. $$k = \hat{\sigma}^2 \frac{\Sigma\lambda_i^2\hat{\alpha}_i^2}{\Sigma\lambda_i^2\hat{\alpha}_i^4}$$

Lawless and Wang $$k = \frac{p\sigma^2}{\Sigma\lambda_i\hat{\alpha}_i^2}$$

Table 2

Correlations, Means, and Standard Deviations of Variables in the Extended Malmö Model of Ability and Achievement (N = 835 Males)[a]

VARIABLE		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Mean	SD
X_1	FATHED		.484	.593	<u>-0.084</u>	.238	.476	.300	.330	.316	.134	2.641	1.206
X_2	FAMINC	734		.830	-0.113	.212	.430	.292	.307	.374	.226	3.600	1.587
X_3	FATHOC 38	764	753		-0.230	.307	.510	.383	.357	.380	.241	2.100	1.00
X_4	FAMSZ	760	753	785		-0.190	-0.241	-0.233	<u>-0.118</u>	<u>-0.099</u>	-0.086	2.559	1.556
X_5	IQ 38	788	752	788	783		.408	.751	.352	.370	.312	97.738	16.021
X_6	EDUC	580	553	578	575	599		.567	.584	.515	.313	1.839	0.950
X_7	IQ 48	629	600	626	623	653	499		.526	.441	.340	97.577	16.474
X_8	OCC 71	548	522	544	544	566	497	461		.507	.357	3.840	1.360
X_9	INC 71	741	705	739	735	773	594	629	564		.720	40.831	29.435
X_{10}	LOGINC[b]	741	705	739	735	773	594	629	564	777		3.423	0.965

(a) Correlation coefficients are above the diagonal. The figures below the diagonal represent the case base for each correlation coefficient. All coefficients are significantly different from zero at the $p \leq .001$ level except for the three underlined coefficients which are significant at the $p < .01$, but $p > .001$. The key to the mnemonics used is as follows: FATHED = father's educational level; FAMINC = parents' income; FATHOC 38 = father's occupation in 1938 at the time the respondent was in the third grade of the Malmö school system; FAMZ = number of siblings including foster children in respondent's family; IQ 38 = respondent's mental ability in 1938 at age 10; EDUC = years and type of schooling completed by respondent; IQ 48 = respondent's mental ability measured at the time of induction into the Swedish armed forces in 1948; OCC 71 = occupational status in 1971; INC 71 = pretax income (all sources) 1971.

(b) The case base for LOGINC and INC 71 is the same (N = 777).

Hoerl, Kennard & Baldwin $k = \frac{p\sigma^2}{\Sigma\hat{\alpha}_i^2}$

The path coefficients for the different stages of the human capital model are tabulated in Table 3 to Table 7, and the characteristics and performance indices of each method are summarized in Tables 8 through 12.

Tables 3 - 12 About Here

Discussion

A. The Condition of the Data Set: For ease of comparison the measure of the degree of multicollinearity for both V_{max} and D_{max}, the reduction of variance, and some other performance indices of simple ridge regression based on the variance normalization criterion (abbreviated to SRR(k_N)) for each stage of the Malmö model were retabulated together in Table 13.

Table 13 About Here

From the D-measure of multicollinearity, we know that each stage has about a 0.8 degree of multicollinearity. From the V_{max}-measure, we see that the maximum variance inflation factor in each stage is about 4, and if we inspect the VIF-matrix for each stage given in appendix B, we see that the third variable FATHOCC is the one that always has the highest VIF. Therefore, the multicollinearity problem in this model centers mainly on the FATHOCC(X_3) variable. From the correlation

Table 3

Path Coefficients and their t-values* (in parenthesis)
for the First Stage.

Dependent Variable: IQ38

	X_1 FATHED	X_2 FAMINC	X_3 FATHOCC	X_4 FAMSZ				
OLS	0.095 (2.34)	-0.104 (1.77)	0.309 (4.73)	-0.123 (3.62)				
Normalization	0.100 (3.17)	-0.017 (0.507)	0.205 (5.80)	-0.119 (4.14)				
Kasarda and Shih	0.097 (2.47)	-0.087 (1.61)	0.289 (4.85)	-0.123 (3.72)				
Hocking, Speed and Lynn	0.098 (2.56)	-0.077 (1.51)	0.277 (4.94)	-0.123 (3.78)				
Hoerl, Kennard and Baldwin	0.098 (2.59)	-0.073 (1.46)	0.272 (4.99)	-0.123 (3.80)				
Lawless and Wang	0.099 (2.62)	-0.070 (1.43)	0.269 (5.01)	-0.1234 (3.81)				
Vincd	0.088 (4.19)	0.030 (1.59)	0.135 (7.43)	-0.096 (4.56)				

*T-values are approximately correct where the sample size is large. This is especially the case when k-values are also low. Both conditions are met in all five stages of the Malmö model. This is also applicable to Table 4-7.

Table 4

Path Coefficients and their t-values (in parenthesis)
for the Second Stage.

Dependent Variable: EDUC

	X_1 FATHED	X_2 FAMINC	X_3 FATHOCC	X_4 FAMSZ	X_5 IQ38			
OLS	0.257 (7.55)	0.093 (1.88)	0.176 (3.16)	-0.121 (4.23)	0.250 (8.62)			
Normalization	0.231 (8.57)	0.103 (3.54)	0.169 (5.55)	-0.113 (4.61)	0.227 (9.12)			
Kasarda and Shih	0.234 (8.47)	0.103 (3.35)	0.169 (5.28)	-0.114 (4.57)	0.229 (9.07)			
Hocking, Speed and Lynn	0.256 (7.60)	0.094 (1.95)	0.175 (3.26)	-0.120 (4.25)	0.249 (8.64)			
Hoerl, Kennard and Baldwin	0.253 (7.73)	0.095 (2.15)	0.174 (3.54)	-0.120 (4.30)	0.246 (8.70)			
Lawless and Wang	0.255 (7.64)	0.094 (2.01)	0.175 (3.34)	-0.120 (4.26)	0.248 (8.66)			
Vinod	0.185 (10.41)	0.108 (6.67)	0.153 (9.98)	-0.094 (5.28)	0.181 (10.11)			

Table 5

Path Coefficients and their t-values (in parenthesis) for the Third Stage.

Dependent Variable: IQ48

	X_1 FATHED	X_2 FAMINC	X_3 FATHOCC	X_4 FAMSZ	X_5 IQ38	X_6 EDUC		
OLS	-0.015 (0.581)	0.004 (0.106)	0.043 (1.02)	-0.037 (1.69)	0.615 (26.85)	0.291 (11.09)		
Normalization	0.005 (0.248)	0.012 (0.534)	0.052 (2.30)	-0.047 (2.53)	0.537 (28.14)	0.264 (12.83)		
Kasarda and Shih	-0.010 (0.396)	0.006 (0.179)	0.046 (1.311)	-0.040 (1.91)	0.595 (27.21)	0.284 (11.52)		
Hocking, Speed and Lynn	-0.015 (0.526)	0.004 (0.113)	0.044 (1.05)	-0.037 (1.72)	0.613 (26.90)	0.290 (11.14)		
Hoerl, Kennard and Baldwin	-0.014 (0.551)	0.004 (0.117)	0.044 (1.05)	-0.037 (1.73)	0.612 (26.91)	0.290 (11.16)		
Lawless and Wang	-0.015 (0.559)	0.004 (0.114)	0.044 (1.06)	-0.037 (1.72)	0.612 (26.90)	0.290 (11.14)		
Vinod	0.034 (2.75)	0.029 (2.61)	0.064 (6.12)	-0.055 (4.41)	0.380 (30.46)	0.206 (16.82)		

Table 6

Path Coefficients and their t-values (in parenthesis) for the Fourth Stage.

Dependent Variable: OCC71

	X_1 FATHED	X_2 FAMINC	X_3 FATHOCC	X_4 FAMSZ	X_5 IQ38	X_6 EDUC	X_7 IQ48	
OLS	0.041 (1.18)	0.032 (0.66)	0.008 (0.15)	0.052 (1.83)	-0.082 (-2.01)	0.393 (10.75)	0.352 (7.80)	
Normalization	0.052 (2.00)	0.035 (1.30)	0.028 (1.00)	0.034 (1.41)	-0.008 (0.314)	0.334 (12.41)	0.271 (9.71)	
Kasarda and Shih	0.044 (1.35)	0.032 (0.78)	0.013 (0.29)	0.048 (1.75)	-0.061 (1.66)	0.380 (11.16)	0.329 (8.22)	
Hocking, Speed and Lynn	0.042 (1.23)	0.032 (0.69)	0.0096 (0.18)	0.051 (1.81)	-0.076 (1.9[illegible])	0.0390 (10.87)	0.346 (7.91)	
Hoerl, Kennard and Baldwin	0.043 (1.28)	0.032 (0.72)	0.011 (0.23)	0.050 (1.78)	-0.070 (1.81)	0.386 (10.99)	0.339 (8.04)	
Lawless and Wang	0.042 (1.25)	0.032 (0.704)	0.010 (0.20)	0.050 (1.80)	-0.073 (1.87)	0.388 (10.93)	0.342 (7.97)	
Vinod	0.061 (3.77)	0.043 (2.93)	0.048 (3.41)	0.009 (0.53)	0.046 (2.94)	0.240 (14.86)	0.191 (12.75)	

Table 7

Path Coefficients and their t-values (in parenthesis) for the Fifth Stage.

Dependent Variable: INC71

	X_1 FATHED	X_2 FAMINC	X_3 FATHOCC	X_4 FAMSZ	X_5 IQ38	X_6 EDUC	X_7 IQ48	X_8 OCC71
OLS	0.0149 (.418)	0.188 (3.74)	-0.042 (-.744)	0.034 (1.15)	0.130 (3.09)	0.227 (5.64)	0.041 (0.847)	0.264 (7.39)
Normalization	0.0247 (0.919)	0.131 (4.66)	0.0192 (0.662)	0.0285 (1.17)	0.103 (3.77)	0.195 (6.84)	0.074 (2.56)	0.229 (8.51)
Kasarda and Shih	0.017 (0.525)	0.164 (4.04)	-0.016 (-0.349)	0.033 (1.20)	0.119 (3.28)	0.216 (5.98)	0.054 (1.34)	0.253 (7.73)
Hocking, Speed and Lynn	0.015 (0.436)	0.182 (3.81)	-0.035 (-0.660)	0.034 (1.16)	0.127 (3.14)	0.224 (5.72)	0.044 (0.952)	0.261 (7.46)
Hoerl, Kennard and Baldwin	0.016 (0.500)	0.168 (3.98)	-0.020 (0.429)	0.033 (1.19)	0.121 (3.24)	0.218 (5.91)	0.052 (1.24)	0.255 (7.66)
Lawless and Wang	0.016 (0.453)	0.177 (3.86)	-0.030 (0.592)	0.034 (1.17)	0.125 (3.17)	0.222 (5.77)	0.047 (1.04)	0.260 (7.52)
Vinod	0.040 (2.39)	0.094 (6.13)	0.050 (3.47)	0.0130 (0.755)	0.085 (5.31)	0.152 (9.30)	0.089 (5.78)	0.176 (10.48)

Table 8

The Characteristics and Performance Indices of OLS Regression and SRR Procedures.

Dependent Variable: IQ38

		k	V_{max}	D_{max}	E.S. Ratio	RVAR	ΔRSS	α	R^2	$\hat{Eft}$
OLS		0	4.06	0.799	1.00	0	0	1.00	0.119	NA
Normalization	k_N	0.146	1.17	0.110	0.069	59.9%	0.38%	0.536	0.096	0.341
Kasarda & Shih	k_S	0.016	3.35	0.744	0.665	13.8%	0.013%	0.999	0.115	2.09
Hocking, Speed and Lynn	k_H	0.027	2.96	0.699	0.509	21.6%	0.033%	0.991	0.113	1.27
Hoerl, Kennard and Baldwin	k_B	0.033	2.81	0.678	0.455	24.7%	0.044%	0.985	0.112	1.09
Lawless & Wang	k_W	0.036	2.71	0.664	0.424	26.5%	0.052%	0.980	0.111	1.00
Vinod	k_V	0.538	0.421	-0.334	0.0037	85.1%	1.38%	0.022	0.068	0.183

Where k = biasing parameter
V_{max} = maximum variance inflation factor
D_{max} = maximum relative degree of multicollinearity
E.S. Ratio = empirical sensitivity ratio
RVAR = relative reduction in variance
ΔRSS = inflation in residual sum of squares
α = α-acceptance level
R^2 = multiple R square
$\hat{Eft}$ = the OLS estimate of effectiveness index

Table 9

The Characteristics and Performance Indices of OLS Regression and SRR Procedures.

Dependent Variable: EDUC

		k	V_{max}	D_{max}	E.S. Ratio	RVAR	ΔRSS	α	R^2	$\hat{Eft}$
OLS		0	4.17	0.805	1.00	0	0	1.00	0.383	NA
Normalization	k_N	0.138	1.24	0.153	0.303	55.6%	0.28%	0.806	0.338	3.16
Kasarda & Shih	k_S	0.121	1.39	0.235	0.331	52.4%	0.21%	0.873	0.343	3.74
Hocking, Speed and Lynn	k_H	0.005	3.89	0.788	0.923	4.79%	0.0005%	1.00	0.381	127.2
Hoerl, Kennard and Baldwin	k_B	-0.020	3.26	0.735	0.756	15.9%	0.007%	1.00	0.376	31.3
Lawless & Wang	k_W	0.010	3.69	0.774	0.868	8.27%	0.002%	1.00	0.380	69.3
Vinod	k_V	0.528	0.430	-0.330	0.096	82.7%	2.60%	0.001	0.250	0.601

Where k = biasing parameter
V_{max} = maximum variance inflation factor
D_{max} = maximum relative degree of multicollinearity
E.S. Ratio = empirical sensitivity ratio

RVAR = relative reduction in variance
ΔRSS = inflation in residual sum of squares
α = α-acceptance level
$\hat{R}^2$ = multiple R square
Eft = the OLS estimate of effectiveness index

Table 10

The Characteristics and Performance Indices of OLS Regression and SRR Procedures.

Dependent Variable: IQ48

		k	V_{max}	D_{max}	E.S. Ratio	RVAR	ΔRSS	α	R^2	$\hat{Eft}$
OLS		0	4.22	0.808	1.00	0	0	1.00	0.648	NA
Normalization	k_N	0.145	1.22	0.138	0.284	54.5%	1.97%	0.013	0.536	0.410
Kasarda & Shih	k_S	0.033	2.89	0.690	0.691	21.9%	0.12%	0.983	0.619	2.48
Hocking, Speed and Lynn	k_H	0.003	4.04	0.798	0.959	28.4%	0.001%	1.00	0.645	28.1
Hoerl, Kennard and Baldwin	k_B	0.005	3.94	0.791	0.936	4.46%	0.004%	1.00	0.643	17.35
Lawless & Wang	k_W	0.004	4.01	0.796	0.953	3.25%	0.002%	1.00	0.645	24.4
Vinod	k_V	0.641	0.368	-0.359	0.063	84.9%	18.6%	0.00	0.332	0.071

Where k = biasing parameter

V_{max} = maximum variance inflation factor

D_{max} = maximum relative degree of multicollinearity

E.S. Ratio = empirical sensitivity ratio

RVAR = relative reduction in variance

ΔRSS = inflation in residual sum of squares

α = α-acceptance level

$\hat{R}^2$ = multiple R square

Eft = the OLS estimate of effectiveness index

Table 11

The Characteristics and Performance Indices of
OLS Regression and SRR Procedures.

Dependent Variable: OCC 71

		k	V_{max}	D_{max}	E.S. Ratio	RVAR	ΔRSS	α	R^2	$\hat{Eft}$
OLS		0	4.22	0.808	1.00	0	0	1.00	0.406	NA
Normalization	k_N	0.160	1.12	0.078	0.114	59.7%	1.12%	0.233	0.338	0.453
Kasarda & Shih	k_S	0.031	2.96	0.699	0.490	21.6%	0.071%	0.999	0.389	2.29
Hocking, Speed and Lynn	k_H	0.008	3.81	0.782	0.813	6.77%	0.006%	1.00	0.401	8.78
Hoerl, Kennard and Baldwin	k_B	0.017	3.43	0.751	0.658	13.3%	0.024%	1.00	0.396	4.15
Lawless & Wang	k_W	0.012	3.63	0.769	0.734	9.86%	0.012%	1.00	0.399	5.82
Vinod	k_V	0.611	0.384	-0.352	0.068	86.5%	6.01%	0.00	0.243	0.155

Where k = biasing parameter
V_{max} = maximum variance inflation factor
D_{max} = maximum relative degree of multicollinearity
E.S. Ratio = empirical sensitivity ratio

RVAR = relative reduction in variance
ΔRSS = inflation in residual sum of squares
α = α-acceptance level
R^2 = multiple R square
$\hat{Eft}$ = the OLS estimate of effectiveness index

Table 12

The Characteristics and Performance Indices of OLS Regression & SRR Procedures.

Dependent Variable: INC 71

		k	V_{max}	D_{max}	E.S. Ratio	RVAR	ΔRSS	α	R^2	$\hat{Eft}$
OLS		0	4.22	0.808	1.00	0	0	1.00	0.372	NA
Normalization	k_N	0.162	1.11	0.072	0.059	59.0%	0.53%	0.818	0.328	0.780
Kasarda & Shih	k_S	0.043	2.62	0.649	0.367	27.0%	0.06%	1.00	0.358	2.20
Hocking, Speed and Lynn	k_H	0.009	3.78	0.780	0.792	7.1%	0.004%	1.00	0.369	9.36
Hoerl, Kennard and Baldwin	k_B	0.034	2.86	0.687	0.442	22.6%	0.046%	1.00	0.361	2.67
Lawless & Wang	k_W	0.016	3.47	0.755	0.661	12.2%	0.012%	1.00	0.366	5.29
Vinod	k_V	0.611	0.38	-0.352	0.004	86.1%	2.9%	0.003	0.253	0.355

Where k = biasing parameter

V_{max} = maximum variance inflation factor

D_{max} = maximum relative degree of multicollinearity

E.S. Ratio = empirical sensitivity ratio

RVAR = relative reduction in variance

ΔRSS = inflation in residual sum of squares

α = α-acceptance level

R^2 = multiple R square

$\hat{Eft}$ = the OLS estimate of effectiveness index

Table 13

The Characteristics and Performance Indices of OLS Regression and SRR(k_N) Procedure for Different Stages.

	OLS			SRR(k_N)							
Dependent Variable	V_{max}	D_{max}	R^2	k	V_{max}	D_{max}	RVAR	ΔRSS	R^2	$\hat{Eft}$	$\hat{Eft}$*
1st Stage IQ38	4.06	0.799	0.119	0.146	1.17	0.110	59.9%	0.38%	0.096	0.34	1.27
2nd Stage EDUC	4.17	0.805	0.383	0.138	1.24	0.153	55.6%	0.21%	0.338	3.16	4.30
3rd Stage IQ48	4.22	0.808	0.648	0.145	1.22	0.138	54.5%	1.97%	0.536	0.41	0.556
4th Stage OCC 71	4.22	0.808	0.406	0.160	1.12	0.078	59.7%	1.12%	0.338	0.45	0.978
5th Stage INC 71	4.22	0.808	0.372	0.162	1.11	0.072	59.0%	0.53%	0.328	0.78	2.40

matrix (Table 2), it is clear that the multicollinearity is attributable to the high correlation between FATHOCC(X_3) and FAMINC(X_2). (Note that high correlation is a sufficient but not necessary condition for severe multicollinearity). Therefore, it may be wise to collapse these two variables in the second phase of the data analysis in order to reduce the multicollinearity problem. At least it is of heuristic interest.

B. <u>The Change Produced by SRR(k_N)</u>: From the second part of Table 13, we can clearly see that, all the harmful effects of multicollinearity at each stage have been greatly reduced by simple ridge regression based on the normalization criterion. The maximum VIFs at each stage have been reduced to about 1.2; the relative degree of multicollinearity at each stage has dropped from about 0.8 to about 0.1; the variances at each stage have been reduced by 55-60 percent; and the sensitivity to fluctuation at each stage due to sampling error has dropped to very small values (by a factor of 48 for the last stage to a factor of 1000 for the second stage). All these indices illustrate the "gain" produced by SRR(k_N). Unfortunately, however, we will never be able to find out the "trade-off" in bias as was pointed out in chapter 3. We cannot even be sure whether the MSE has been reduced. If we look at the OLS estimate of the effectiveness index, $\hat{Eft}$, only the second stage has a value larger than unity; that is, the reduction in variance is greater than the total square bias, or stated differently, the MSE has been reduced. For the remaining stages, the $\hat{Eft}$'s are less than one. However, as was pointed out in chapter 3 $\hat{Eft} > 1$ is a sufficient but not necessary condition for proving the MSE has been reduced. Therefore, except for the second stage, we cannot be positive that we have obtained better estimates. Based on the theoretical argument in chapter 4, if we assumed that $SRR(k_N)$ produced more accurate estimates, then the ridge estimate of the effectiveness index, $\hat{Eft}^*$, for all stages would have an average greater than unity. Although there is no clear cut index ensuring that the $SRR(k_N)$ procedure has really improved the estimates, from the small amount of inflation in the residual sum of squares and the large reduction in variance; intuitively, it is believed that, the $SRR(k_N)$ procedure has produced better estimates.

C. The Change in Path Coefficients Produced by $SRR(k_N)$: The path coefficients produced by OLS, $SRR(k_N)$ and other methods were tabulated in Table 3 to 7 for stage 1 to stage 5 respectively. If we compare the path coefficients produced by OLS and $SRR(k_N)$ procedures, we would observe the following.

1. In the first stage (Table 3, IQ38 as dependent variable), the negative effect of $FAMINC(X_2)$ on IQ38 became insignificant at the 0.05 level.
2. In the second stage (Table 4, EDUC as dependent variable), no dramatic change in path coefficients resulted from $SRR(k_N)$.
3. In the third stage (Table 5, IQ48 as dependent variable), the effect of $FATHOCC(X_3)$ on IQ48 became significant at about the 0.01 level.

4. In the fourth stage (Table 6, OCC71 as dependent variable), the effect of IQ38 on OCC71 became insignificant.
5. In the last stage (Table 7, INC71 as dependent variable), the effect of IQ48 on INC71 became significant at about the 0.005 level.
6. At each stage, there was no major change in the rank order of magnitude of the estimated path coefficients. However, the overall SRR(k_N) path estimates were significantly different from those of the OLS estimates. This can be seen from the so called "acceptance level" or "associated probability" which is 1.00 for OLS. The deviation from unity shows the level of deviation from the OLS estimates (Obenchain 1978, McCabe 1978).

Based on the above changes, those variables that are not significant at the 0.05 level were dropped and the model was reanalyzed by using both OLS regression and SRR(k_N). The results are summarized in the path diagrams depicted in Figures 5A and B. Figure 5A gives the OLS regression results and Figure 5B presents the results for simple ridge regression based on the normalization criterion.

Figure 5A & 5B About Here

D. <u>The Comparison of Different Types of SRR</u>: The characteristics and performance indices of OLS regression and different types of SRR for different stages were tabulated in

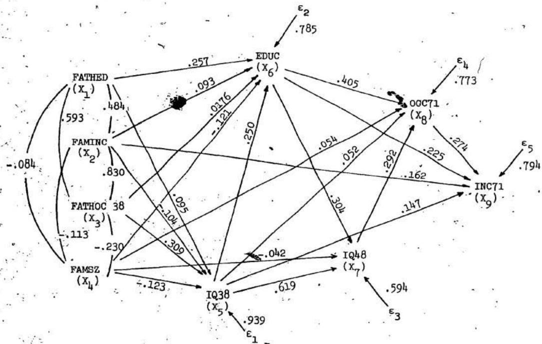


Figure 5A: The Path Diagram Obtained by OLS Regression

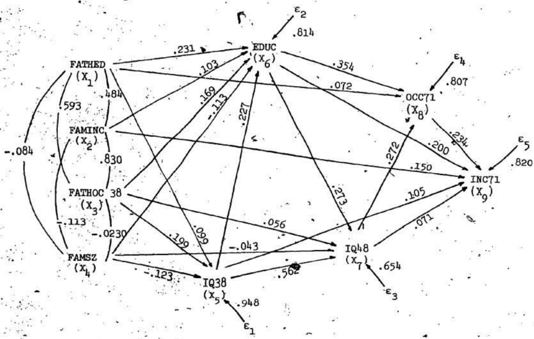


Figure 5B: The Path Diagram Obtained by SRR(k_N)

Table 8 to Table 12. The discussion here will follow the parameters or indices listed in the tables.

1. The bias parameter k: It is obvious from Tables 8 - 12, that Vinod's method produced the largest k, that the variance normalization method produced moderate k, and all other SRR methods generally produced much smaller k.

2. The maximum VIF (V_{max}): The maximum variance inflation factor is a linear measure of the severity of multicollinearity in the data set. For perfectly orthogonal data it is equal to one. From column 2 in tables 8 - 12, it is obvious that the variance normalization method is always the more appropriate one. Vinod's method always produced a V_{max} far less than unity and hence might produce too large a bias, while the rest of the SRR procedures have not produced enough reduction in VIF; that is, have not minimized the harmful effects of multicollinearity.

3. The D-measure of multicollinearity (D_{max}): As has been pointed out in chapter 2, the D-measure of multicollinearity has a range of 0 to 1; $D = D_{max} = 0$ for no multicollinearity, and $D = D_{max} = 1$ for perfect multicollinearity. From column 3 in tables 8 - 12, we see that the variance normalization method always produces a D_{max} close to zero as desired; Vinod's method always resulted in D_{max} smaller than zero; and the rest of the SRR procedures always had higher degrees of multicollinearity.

4. The Empirical Sensitivity Ratio (E.S. Ratio): The empirical sensitivity ratio may be measured by the ratio of the sum of squares of the fluctuation of estimated ridge regression coefficients (SSF*) over those of OLS regression (SSF), that is

$$\text{E. S. Ratio} = \frac{SSF^*}{SSF}$$

$$= \frac{\Sigma(\hat{\beta}_i^* - \tilde{\beta}_i^*)^2}{\Sigma(\hat{\beta}_i - \tilde{\beta}_i)^2}$$

where $\tilde{\beta}_i$ and $\tilde{\beta}_i^*$ represent the ith perturbed OLS regression coefficient and that of the ridge regression coefficient respectively; and where, the perturbed coefficients are obtained by introducihg a small perturbation Δr to the largest correlation coefficient in the correlation matrix $X'X$. The amount of perturbation used in the analysis is $|\Delta r| = 0.01$, which is a reasonable approximation to the precision of a correlation coefficient in most measurement.

From the value of E.S. Ratio in tables 8 - 12, it is obvious that the variance normalization method has very low E.S. Ratio values compared to most of the SRR procedures with the exception of Vinod's method.

5. The Relative Reduction in Variance (RVAR): The relative reduction in variance is a measure of the percentage reduction of the total variance in

the parameter estimates generated by the ridge procedure over the OLS regression procedure. By definition it can be expressed as

$$RVAR = \frac{tr(X'X)^{-1} - tr(VIF)}{tr(X'X)^{-1}}$$

$$= 1 - \frac{tr(VIF)}{tr(X'X)^{-1}}$$

From column 5 in tables 8-12, it is obvious that the normalization method always produces a large amount of reduction (55-60%); and although it is not the highest, it is usually higher than the other SRR procedures included in this study.

6. The Inflation in the Residual Sum of Squares: The OLS regression is based on the minimum residual sum of squares criterion. The use of a different criterion such as MSE would produce a larger residual sum of squares. From column 6 in the tables we see that except for Vinod's method, the variance normalization method and the other SRR procedures produce trivial amounts of inflation in the residual sum of squares.

7. The acceptance level: The acceptance level of a ridge estimate, $\hat{\beta}^*$, is defined by McCabe (1978) as the significance level of the F-ratio.

$$F_{p,n-p,(1-\alpha)} = \frac{(\hat{\beta} - \hat{\beta}^*)' X'X (\hat{\beta} - \hat{\beta}^*)}{p\sigma^2}$$

A ridge estimate is called "α- acceptable" if the estimate is in the $(1 - \alpha)$ 100% confidence region.

According to McCabe (1978) and Obenchain (1977) it is desirable for an estimate to have a high acceptance level. Although the author does not agree with their argument, the α acceptance levels at each stage, except stage 3, for the variance normalization method are acceptably high.

8. The multiple R square: From column 9 in tables 8 - 12, it is obvious that, except for Vinod's and the variance normalization method, the remaining SRR procedures still maintained a high R^2 (relative to that of OLS regression). For the normalization method, the reduction at each stage, although not as large as those for Vinod's method, were fairly large. This reduction in R^2 might make some people hesitate to use ridge regression due to their misunderstanding over the importance of R^2. They might erroneously believe that the reduction in R^2 indicates that the ridge procedure provides a poorer fit, with subsequent reduction in the predictive power of the regression model. This conclusion is definitely not correct for at least the following three reasons. Firstly, even if we use OLS regression, whether we can use R^2 as a measure of goodness of fit is doubtful (Barrett, 1974; Pindyck & Rubinfeld, 1976, pp. 61). Secondly, the minimum residual sum of squares (equivalent to maximizing R^2) criterion has been shown not to perform well empirically. Thirdly,

it has been pointed out by many researchers, and has been mathematically proven in this study (chapter 2) that while to have large R^2 is desirable, it is not sufficient to ensure that model will have high predictive power. As has been shown, the predictive power of a regression model is very sensitive to the variance of the estimates or the degree of the multicollinearity of the data set. If we compare the reduction in R^2 and the amount of variance reduced, we would see that (from Eq. 3.33) if the total square bias is not too large, the ridge estimates would have much better predictive power. Judging the reduction in R^2, and the reduction in variance, that the bias parameter k introduced, we would conclude that the normalization method is more conservative than Vinod's method, and has a much better chance of performing better in prediction.

9. The OLS estimate of the effectiveness index $\widehat{Eft}$:
The effectiveness index is defined as the ratio of reduction in total variance over the total square bias. Its true value cannot be calculated due to its dependence on unknown population parameters. As has been pointed out in chapter 3, its OLS estimate gives us a conservative (larger) value. Therefore, having $\widehat{Eft} > 1$ is a sufficient but not a necessary condition for proving that the estimate

has lower MSE. That is, it is a better estimate in this sense. From column 9 in tables 8 - 12, it is obvious that only Kasarda and Shih's, Hocking et. al.'s, HKB's and Lawless and Wang's methods have $\hat{E}ft$ greater than unity; that is, can assure us the MSE is not larger than that of OLS. However, as has been pointed out previously, these methods actually do not produce significant differences for the better in all respects as compared to OLS estimates. Stated differently, these methods have not performed well in the reduction of the harmful effects of multicollinearity.

E. The Ridge Trace: In this section, only the ridge trace for the last stage in the Malmö model was plotted. The purpose of the trace was not to determine the optimal k due to the subjectiveness of the ridge trace method, instead it is used to display the characteristics of the data set, and to compare the k's obtained from different simple ridge regression procedures. The ridge trace for the last stage is given in Figure 6.

Figure 6 About Here

From the location of the k's on the ridge trace, it is obvious that the k estimated by Hocking et. al. (k_H), Kasarda and Shih (k_S), Hoerl, Kennard & Baldwin (k_B) and Lawless & Wang (k_W) are too low as compared to the guidelines for determining k given by Hoerl and Kennard (see chapter 3, pp. 38-39).

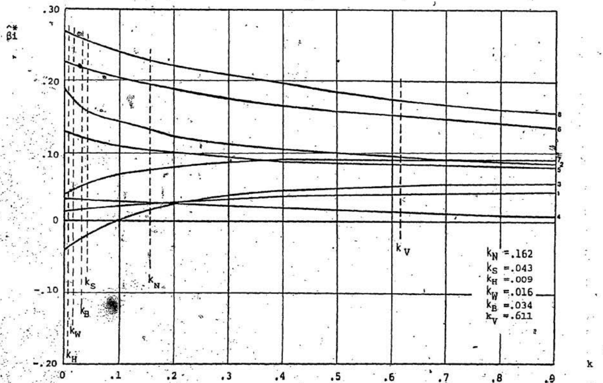


Figure 6: The Ridge Trace for the Last Stage in Malmö Model

where k_H = estimated k by using Hocking et.al.'s method

k_B = estimated k by using Hoerl, Kennard and Baldwin's method

k_N = estimated k by using normalization criterion

k_W = estimated k by using Lawless and Wang's method

k_S = estimated k by using Kasarda and Shih's method

k_V = estimated k by using Vinod's method

Vinod's method (k_V) definitely gives too high an estimate, and it seems that the variance normalization method (k_N) is the only one that satisfies the guidelines.

From the comparison of the k values obtained by different methods for various stages (see tables 8 – 12), it is obvious that, if the ridge traces were plotted, similar results and conclusions as that for the last stage would be reached. To further support this conclusion, the ridge trace with various k's for the 10-factor problem described by Hoerl and Kennard (1970b), and the 5-factor problem given by Price (1977) are presented in appendix C. From the comparison of all these traces, it is clear that the variance normalization method seems to consistently produce a biasing parameter which satisfies Hoerl and Kennard's guidelines for obtaining optimal k. This evidence indicates that the variance normalization method is superior to the other methods included in this study.

Summary and Conclusion

Based on the analysis of a career achievement model by using the Malmö data set, we observed the following important facts about simple ridge regression based on the variance normalization criterion.

1. It produced, compared to other SRR procedures, the maximum reduction in the harmful effects of multicollinearity.
2. It gave much more stable estimates of the regression coefficients than OLS regression or any other SRR procedures included in the analysis.

3. It produced only a small amount of inflation in the residual sum of squares.

4. It produced the only "k" that satisfied Hoerl and Kennard's guidelines for selecting optimal k.

On the basis of these facts, together with the facts that with the variance normalization method,

(i) the k can be calculated accurately,

(ii) it is nonstochastic and equal to zero for an orthogonal data set as required, and

(iii) it consistently satisfies the empirical requirements of optimal k suggested by Hoerl and Kennard,

the variance normalization method is likely to be superior to the other methods examined in this study.

CHAPTER VI

CONCLUSION AND FURTHER RESEARCH

Conclusion

It was stated in chapter I that this study has the following purposes: (1) to demonstrate the superiority of simple ridge regression over OLS regression through theoretical argument and empirical example, (2) to modify ridge regression through use of the variance normalization criterion in order to achieve more satisfactory empirical results, and (3) to demonstrate the superiority of simple ridge regression based on the variance normalization criterion over those ridge regression estimates based on minimum mean square errors. The theoretical discussion in chapter II, III and IV and the empirical study in chapter V have constituted efforts to fulfill these purposes.

In the chapter III discussion of the properties of ridge regression it was shown that in ridge regression the total variance of the estimated regression coefficients was greatly reduced by the introduction of a small bias in the estimates, and that the procedure gave a better estimate as long as the amount of reduction in the total variance was greater than the total squared bias introduced. It has been shown that under this condition ridge estimates would have smaller total variance; would be more reliable; would be less sensitive to sampling error or model misspecification error; and would have better predictive power.

In the theoretical discussion in chapter III and chapter

IV we showed that ridge regression based on the minimum MSE criterion has several limitations: (1) the theoretical value of optimal k, which gives the minimum MSE, is stochastic though it should be non-stochastic; (2) the true optimal k depends on population parameters and cannot be obtained; (3) even for an orthogonal data set there is an optimal k which produces minimum MSE, and (4) the bias produced by the ridge procedure, which is the "trade-off" required in order to achieve all the desirable properties of ridge regression, cannot be accurately estimated. We have also pointed out that the origin of these limitations lies in application of the MSE criterion.

Further, through the analysis of the source of variance produced by OLS regression, we have pointed out that ridge regression based on the minimum MSE is multifunctional. Thus, it may reduce the variance from three sources: that originating from the random errors in the designed model; that due to small sample size; and the variance inflated by the multicollinearity problem. From this analysis we have developed a "unifunctional" ridge regression procedure designed solely to "normalize" the variance inflated by the multicollinearity problem. The criterion, called the variance normalization criterion, avoids the first three limitations; that is, through use of the variance normalization criterion the resultant k is nonstochastic, it can be calculated exactly, and for orthogonal data sets it is equal to zero.

Through formulation and estimation of a structural equation model dealing with a problem in human capital theory, and through use of the longitudinal Malmö data set, we have provided support for the theoretical arguments presented in chapter III and chapter IV; we have also shown that the results at each stage of the model, obtained by ridge regression based on the variance normalization criterion, drastically reduced the total variance. Thus, the results were far less sensitive to sampling error, compared to those generated by OLS regression or most of the MSE ridge procedures included in this study.

As pointed out earlier, with the normalization criterion we have avoided three out of the four dilemmas of ridge regression. With regard to the fourth dilemma; namely, the estimation of the total squared bias, due to its dependence on population parameters, no solution is possible.

Although all research studies have indicated that theoretically, ridge regression is superior to those alternatives currently under investigation; for example, the principal component method, Stein's shrunken estimator, (see for example Dempster et. al. 1977, Marquardt 1970, Hocking, Speed, and Lynn, 1976), empirically it is a completely different matter. Due to the fact that we cannot estimate the bias accurately, it is impossible to show the performance of any ridge procedure directly and accurately. Therefore, sometimes (when $Eft > 1$ or k is not in the admissible range $0 < k < k_{max}$), we cannot be sure whether the ridge estimates are really better than OLS estimates. However, it has been argued (Marquardt and Snee 1976) that due to the wide range of k which produce smaller MSE, most methods with conservative k would produce better estimates than OLS regression.

From the example studied in chapter V, we have seen that all methods except the normalization one failed to optimize the reduction of the harmful effects of multicollinearity. Although the results are not presented here, more examples have been studied (cf. Bulcock and Lee 1978, Beebe and Bulcock 1978). These studies have shown that the normalization method always produces more acceptable results than any of the others included in this study. If one accepts the underlying assumption of the normalization criterion that k_N is less than k_{max}, the true maximum k, the theoretical argument and the empirical results have clearly indicated that ridge regression using the variance normalization criterion is superior to other ridge regression estimating procedures.

As a final comment it is worth noting that this study is not to be interpreted as implying that ridge regression is a solution to the multicollinearity problem; it is just a procedure of last resort for reducing the ill-effects of the problem. If physically possible, the source of multicollinearity should be located and eliminated. However, this is seldom feasible. Further, the normalization criterion itself does not always perform better than the minimum MSE criterion. It depends on the purpose of the study. If we are interested mainly in the explanation of a phenomenon, which is usually the case in the social sciences; then the normalization criterion is more suitable. However, if the goal lies in forecasting, an estimator with lower variance and larger bias (which may be empirically estimated), may be more appropriate. For this purpose Vinod's method may be more suitable, as it generally produces a much larger biasing parameter k.

Further Research

There are at least two immediate extensions of this study that have to be performed in the near future. First is the Monte Carlo simulation test of the validity of the normalization criterion. As has been pointed out earlier, theoretically any ridge regression is superior to OLS regression if it can be ensured that the reduction in the total variance is greater than the total squared bias introduced by the ridge procedure. In terms of the effectiveness index or k_{max}, this condition is equivalent to Eft > 1, or $k < k_{max}$. Due to the dependence of the total squared bias on population parameters, the "superiority condition" of ridge regression cannot be evaluated accurately. In chapter III, we have shown that a conservative "test" which gives evidence of the "superiority condition" is one in which $\widehat{Eft} > 1$ or $k < \hat{k}_{max}$, where the "hat" indicates that it is based on OLS estimates (see property xiv in chapter III). From the empirical application discussed in chapter V, we have seen that, the ridge procedures which satisfy this conservative "test" are those which do not perform well in the reduction of the harmful effect of the multicollinearity problem. The one that optimizes the reduction -- the normalization method -- fails this conservative "test". The alternative is to apply a Monte Carlo simulation experiment to in chapter II and chapter V, we know that, both variance and covariance are actually inflated by multicollinearity. Therefore, in the generalized normalization ridge regression (GNR), we need to consider the reduction of the inflated covariance as well. Further, ideally, the correlation between the regression coefficients should also be reduced in order to simulate the situation of orthogonal data set.

perform a test on its general validity.

Because of the crudeness of simple ridge regression, where all the diagonal elements are augmented with the same constant k, the simulation test is not likely to be a hundred percent affirmative. However, it is still of heuristic interest to find out how simple ridge regression using the k_N criterion compares to other estimating procedures. It is worth noting that, this "superiority condition" is also a sufficient condition not a necessary condition due to the fact that the OLS estimate itself is empirically biased. The OLS estimate is unbiased only when the model is a true model (Johnston 1972, pp. 169, Draper & Smith 1966, pp. 81), and empirically regression models are usually misspecified. Therefore, the "superiority condition" is one in which one biased estimate is compared to another biased estimate. In this situation even if the reduction in variance is less than the squared bias, we cannot tell which one is closer to the true value, unless both are biased in the same direction. Although we know that a ridge estimate is always negatively biased, we do not know the direction of the bias in OLS estimates produced by the specification error (see Johnston 1972, pp. 169, Eq. 5-102). Therefore, if the simulation test fails, the last alternative is to subject the model to empirical test.

The second important continuation of this study is the development of the generalized ridge regression based on the normalization criterion. In simple ridge regression, only the inflation in the variance is considered. From the discussion

References

1. Allen, D.M. (1974). The Relationship Between Variable Selection and Data Augmentation and Method for Prediction. Technometrics, Vol. 16, 125-127.

2. Barrett, J.P. (1974). The Coefficient of Determination -- Some Limitations. The American Statistician, Vol. 28, 19-20.

3. Beebe, M.J. and Bulcock, J.W. (1978). Cueing Strategies and Basic Skills in Early Reading. Paper presented at the International Reading Association Seventh World Congress on Reading. Hamburg, August, 1978.

4. Blau, P.M. and Duncan, O.D. (1967). The American Occupational Structure. New York: Wiley.

5. Bulcock, J.W., Fägerlind, I. and Emanuelsson, I. (1974). Education and the Socioeconomic Career, U.S. - Swedish Comparisons. Institute for the Study of International Problems in Education. Report 6, May 1974, University of Stockholm.

6. Bulcock, J.W., Fägerlind, I. and Emanuelsson, I. (1975). Education and the Socioeconomic Career in Sweden: An Overview of Recent Research. Paper prepared for the Annual AERA Meeting. Washington, D.C., April 1975.

7. Bulcock, J.W. and Lee, W.F. (1978). Cross-Cultural Differences in Cognitive Processes: Differences in Degree or of Kind? Presented at the Proceedings of Ninth World Congress of Sociology, Uppsala, August, 1978.

8. Coleman, J.S. (1971). Resources for Social Change. New York: Wiley - Interscience.

9. De Wolff, P., and Van Slijpe, A.R.D. (1973). The Relation between Income, Intelligence, Education, and Social Background. European Economic Review 4, 235-264.

10. Dempster, A.P., Schatzoff, M. and Wermuth, N. (1977). A Simulation Study of Alternatives to Ordinary Least Squares. Journal of the American Statistical Association, 72, 77-91.

11. Draper, N. and Smith, H. (1966). Applied Regression Analysis. Wiley.

12. Emanuelsson, I., Fägerlind, I., and Hartman, S. (1973). Vexunutbildning och arbetsförhållanden. En enkätstudie inom malmöundersökningen av 1938 ars tioåringar i 45-årsaldern. Report No. 96. Pedagogiska institutionen vid Lärarhögskolan i Stockholm.

13. Fägerlind, I. (1975). Formal Education and Adult Earnings. Almqvist and Wiksell International, Stockholm.

14. Farrar, D.E. and Glauber, R.R. (1967). Multicollinearity in Regression Analysis. The Problem Revisited. The Review of Economics and Statistics, 92-107.

15. Guilkey, D.K. and Murphy, J.L. (1975). Directed Ridge Regression Techniques in Case of Multicollinearity. Journal of the American Statistical Association, Vol. 70, 769-775.

16. Hallgren, S. (1939). Intelligens och Miljö. En experimentell undersöking an barn i tredje skolaret vid Malmö folkskolor och privata skolor, I - II. Unpublished lic.-thesis, University of Lund.

17. Hause, J.C. (1972). "Earnings Profile: Ability and Schooling." Journal of Political Economy 80, Part II (May): 108-138.

18. ________ (1975). "Ability and Schooling Determinants of Lifetime Earnings, or If You're So Smart, Why Aren't You Rich?" (Revised). In F.T. Juster (Ed.), Education, Income, and Human Behavior. New York: McGraw-Hill.

19. Hemmerle, W.J. (1975). An Explicit Solution for Generalized Ridge Regression. Technometrics, Vol. 17, 309-314.

20. Hemmerle, W.J. and Brantle, T.F. (1978). Explicit and Constrained Generalized Ridge Estimation. Technometrics, Vol. 20, 109-120.

21. Hocking, R.R. (1976). The Analysis and Selection of Variables in Linear Regression. Biometrics, 32, 1-49.

22. Hocking, R.R., Speed, F.M. and Lynn, M.J. (1976). A Class of Biased Estimators in Linear Regression. Technometrics, Vol. 18, 425-437.

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

24. Hoerl, A.E. and Kennard, R.W. (1970b). Ridge Regression: Application to Nonorthogonal Problems. Technometrics, Vol. 12, 69-82.

25. Hoerl, A.E. and Kennard, R.W. (1976). Ridge Regression: Iterative Estimation of the Biasing Parameter. Communications in Statistics - Theoretical Methods, A5(1), 77-88.

26. Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975). Ridge Regression: Some simulation. Communications in Statistics, 4(2), 105-123.

27. Husen, T. (1947). Några data rörande svenska krigsmaktens inskrivningsprov. Lund.

28. ________ (1948a). Konstruktion och standardisering av svenska krigsmaktens inskrivningsprov. 1948 års version. Stockholm: CVB.

29. ________ (1948b). "Fridtjuf Berg och enhetsskolan." Svensk lärartidning. Pedagogiska skrifter 199.

30. ________ (1950). Testresultatens prognosvärde. Stockholm: Almqvist and Wiksell.

31. Husen, R., and Henricson, S.E. (1951). Some principles of Construction Group Intelligence Tests for Adults. A Report on the Construction and Standardization on the Swedish Induction Test (the I-test). Stockholm: Almqvist and Wiksell.

32. Husen, T., with Emanuelsson, I., Fägerlind, I., and Liljefors, R. (1969). Talent, Opportunity and Career. A twenty-six year follow-up of 1500 individuals. Stockholm: Almqvist and Wiksell.

33. Jencks, C., Smith, M., Ackland, H., Bane, M.J., Cohen, D., Gintis, H., Heyns, B. and Michelson, S. (1972). Inequality: A Reassessment of the Effect of Family and Schooling in America. New York: Basic Books.

34. Johnston, J. (1972). Econometric Methods. Second Edition. McGraw-Hill.

35. Kasarda, J.D. and Shih, W.F. (1977). Optimal Bias in Ridge Regression Approaches To Multicollinearity. Sociological Methods and Research, Vol. 5, No. 4, 461-470.

36. Kmenta, J. (1971). Elements of Econometrics. Macmillan.

37. Lawless, J.F. and Wang, P. (1976). A Simulation Study of Ridge and other Regression Estimators. Communications in Statistics - Theoretical Methods, A5(4), 307-323.

38. Lindley, D.V. and Smith, A.F.M. (1972). Bayes Estimates for the Linear Model. Journal of the Royal Statistical Society, Series B, 34, 1-41.

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

40. Marquardt, D.W. and Snee, R.D. (1975). Ridge Regression in Practice. The American Statistician, Vol. 29, 3-20.

41. McCabe, G.P. Jr. (1978). Evaluation of Regression Coefficient Estimates Using α-Acceptability. Technometrics, Vol. 20, 131-139.

42. McDonald, G.C. and Galarneau, D.I. (1975). A Monte Carlo Evaluation of Some Ridge-Type Estimators. Journal of the American Statistical Association, Vol. 70, 407-416.

43. McDonald, G.C. and Schwing, R.C. (1973). Instabilities of Regression Estimates Relating Air Pollution to Mortality. Technometrics, Vol. 15, 463-481.

44. Obenchain, R.L. (1975). Ridge Analysis Following a Preliminary Test of the Shrunken Hypothesis. Technometrics, Vol. 17, 431-445.

45. Obenchain, R.L. (1977). Classical F-tests and Confidence Region for Ridge Regression. Technometrics, Vol. 19, 429-439.

46. Pindyck, R.S. and Rubinfeld, D.L. (1976). Econometric Methods and Economic Forecasts. McGraw-Hill.

47. Price, B. (1977). Ridge Regression: Application to Nonexperimental Data. Psychological Bulletin, Vol. 84, 759-766.

48. Stein, C. (1955). Inadmissability of the Usual Estimator for the Mean of a Multivariate Normal Distribution. Proc. Third Berkely Symp. Math. Statist. Prob., 1, 197-206.

49. Vinod, H.D. (1976). Application of New Ridge Regression Methods to a Study of Bell System Scale Economics. Journal of the American Statistical Association, 71, 835-841.

50. Vinod, J.D. (1978). A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least Squares. The Review of Economics and Statistics, 60 ; 121-131.

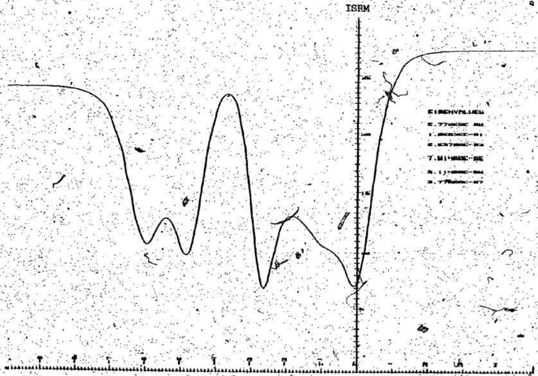
51. Wannacott, R.J. and Wannacott, T.H. (1969). Econometrics. New York: Wiley.

52. Widder, D.V. (1963). Advanced Calculus. Englewood Cliffs, New Jersey: Prentice-Hall.

53. White, K.J. (1978). SHAZAM, Version 2.1, Department of Economics, Rice University. Houston, Texas, U.S.A.

Appendix A.

The multiple minimum of ISRM for the 6-factor example from Vinod (1976)



Appendix B

The VIF-Matrices for the Career Achievement Model 1 (From the First to the Fifth Stage).

First Stage $V_{max} = 4.057$

0.1551D+01	0.6361D-01	-0.9932D+00	-0.9100D-01
0.6361D-01	0.3285D+01	-0.2827D+01	-0.2737D+00
-0.9932D+00	-0.2827D+01	0.4057D+01	0.5303D+00
-0.9100D-01	-0.2737D+00	0.5303D+00	0.1083D+01

Second Stage $V_{max} = 4.166$

0.1561D+01	0.5237D-01	-0.9599D+00	-0.1042D+00	-0.1077D+00
0.5237D-01	0.3297D+01	-0.2864D+01	-0.2591D+00	0.1184D+00
-0.9599D+00	-0.2864D+01	0.4166D+01	0.4873D+00	-0.3508D+00
-0.1042D+00	-0.2591D+00	0.4873D+00	0.1100D+01	0.1392D+00
-0.1077D+00	0.1184D+00	-0.3508D+00	0.1392D+00	0.1135D+01

Third Stage $V_{max} = 4.216$

0.1668D+01	0.9113D-01	-0.8867D+00	-0.1546D+00	-0.3405D-02	-0.4168D+00
0.9113D-01	0.3311D+01	-0.2837D+01	-0.2774D+00	0.1562D+00	-0.1508D+00
-0.8867D+00	-0.2837D+01	0.4216D+01	0.4528D+00	-0.2795D+00	-0.2849D+00
-0.1546D+00	-0.2774D+00	0.4528D+00	0.1124D+01	0.9019D-01	0.1961D+00
-0.3405D-02	0.1562D+00	-0.2795D+00	0.9019D-01	0.1236D+01	-0.4056D+00
-0.4168D+00	-0.1508D+00	-0.2849D+00	0.1961D+00	-0.4056D+00	0.1621D+01

Fourth Stage $V_{max} = 4.221$

0.1669D+01	0.9096D-01	-0.8886D+00	-0.1530D+00	-0.3044D-01	-0.4296D+00	0.4396D-01
0.9096D-01	0.3311D+01	-0.2837D+01	-0.2778D+00	0.1631D+00	-0.1475D+00	-0.1133D+01
-0.8886D+00	-0.2837D+01	0.4221D+01	0.4482D+00	-0.2038D+00	-0.2491D+00	-0.1231D+00
-0.1530D+00	-0.2778D+00	0.4482D+00	0.1128D+01	0.2552D-01	0.1655D+00	0.1052D+00
-0.3044D-01	0.1631D+00	-0.2038D+00	0.2552D-01	0.2311D+01	0.1029D+00	-0.1749D+01
-0.4296D+00	-0.1475D+00	-0.2491D+00	0.1655D+00	0.1029D+00	0.1862D+01	-0.8270D+00
0.4396D-01	-0.1133D+01	-0.1231D+00	0.1052D+00	-0.1749D+01	-0.8270D+00	0.2844D+01

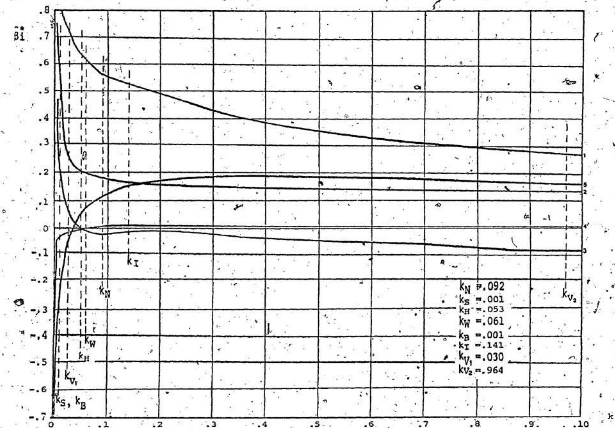
Fifth Stage $V_{max} = 4.221$

0.1671D+01	0.9316D-01	-0.8880D+00	-0.1494D+00	-0.3607D-01	-0.4025D+00	0.6822D-01	-0.6890D-01
0.9316D-01	0.3313D+01	-0.2836D+01	-0.2750D+00	0.1587D+00	-0.1264D+00	0.7619D-02	-0.5382D-01
-0.8880D+00	-0.2836D+01	0.4221D+01	0.4489D+00	-0.2049D+00	-0.2438D+00	-0.1183D+00	-0.1351D-01
-0.1494D+00	-0.2750D+00	0.4489D+00	0.1133D+01	0.1834D-01	0.2000D+00	0.1361D+00	-0.8769D-01
-0.3607D-01	0.1587D+00	-0.2049D+00	0.1834D-01	0.2323D+01	0.4875D-01	-0.1797D+01	0.1378D+00
-0.4025D+00	-0.1264D+00	-0.2438D+00	0.2000D+00	0.4875D-01	0.2122D+01	-0.5941D+00	-0.6615D+00
0.6822D-01	0.7619D-02	-0.1183D+00	0.1361D+00	-0.1797D+01	-0.5941D+00	0.3053D+01	-0.5927D+00
-0.6890D-01	-0.5382D-01	-0.1351D-01	-0.8769D-01	0.1378D+00	-0.6615D+00	-0.5927D+00	0.1683D+01

Appendix C-1

The 5-Factor Employee Satisfaction Problem from Price (1977)

($N = 30$, $V_{max} = 421.8$, $D_{max} = .998$)



Appendix C-2

10-Factor Example from Hoerl and Kennard (1970b)

($N = 36$, $V_{max} = 9.11$, $D_{max} = .952$)