INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

University Microfilms International
300 North Zeeb Road
Ann Arbor, Michigan 48106 USA
St. John's Road, Tyler's Green
High Wycombe, Bucks, England HP10 8HR

77-18,295

BASKIN, David, 1947-
AN EMPIRICAL COMPARISON OF THREE TECHNIQUES FOR THE SELECTION OF A SUBSET OF PREDICTOR VARIABLES.

City University of New York, Ph.D., 1977
Education, psychology

AN EMPIRICAL COMPARISON OF THREE TECHNIQUES FOR THE SELECTION OF A SUBSET OF PREDICTOR VARIABLES

by

David Baskin

A dissertation submitted to the Graduate faculty in Education in partial fulfillment of the requirements for the degree of Doctor of Philosophy, the City University of New York

1977

This manuscript has been read and accepted for the Graduate Faculty in Education in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

3/21/77
date

Chairman of Examining Committee

3/21/77
date

Executive Officer

Dr. Alan Gross

Dr. Damodar Gujarati

Dr. Max Weiner

Supervisory Committee

The City University of New York

Abstract

An Empirical Comparison of Three Techniques for the Selection of a Subset of Predictor Variables

by David Baskin

Advisor: Professor Alan Gross

Attempts to select an optimal subset of predictors from a larger set of predictors are legion; three methods have been discussed and investigated. The Forward Selection procedure was chosen because it is very commonly used by practitioners. The Ridge Selection procedure and Bayesian Selection procedures were chosen since they are more innovative techniques for variable selection and have not been duly examined.

Each selection procedure seeks to minimize the value of some criterion. The Forward Selection procedure seeks to minimize the residual sum of squares in a sample. The Ridge Selection procedure seeks to minimize the mean squared error, i.e., the average squared difference between the value of the population coefficients of the predictors and the sample coefficient estimates. The Bayesian procedure seeks to minimize the total psychometric plus non-psychometric cost with respect to future samples. However, in addition to these criteria, it was discussed how the goal of a selection procedure should more relevantly be directed to two other criteria, namely 1) maximizing the average weight validity, i.e., the correlation between Y and $\hat{Y}$ with respect to the entire population, and 2) minimizing the mean squared error

validity, i.e., the average squared distance between $\underline{B}$ and $\hat{\underline{B}}$, i.e., when $\hat{\underline{B}}$'s derived in a calibration sample are subsequently applied to the population. Therefore, it was necessary to determine which procedures yielded subsets having 1) the lowest average $\bar{R}^2$, 2) the lowest average mean squared error, 3) the lowest average cost, 4) the highest average weight validity, and 5) the highest average mean squared error validity.

The following were the expectations of the results: 1) that the Forward Selection procedure would yield subsets having the largest $\bar{R}^2$ since this procedure seeks to maximize the value of this criterion in selection subsets, 2) that the Ridge procedure would yield subsets having the smallest M.S.E. since the procedure attempts to minimize the value of this criterion, 3) that the Bayesian procedure would yield subsets having the smallest cost since the procedure attempts to minimize the value of this criterion, 4) that, with regard to the average weight validity and average mean squared error validity indices, the Ridge procedure would yield higher index validities since this procedure, which minimizes the M.S.E., would yield better estimates of population parameters, especially when the intercorrelation among predictors is high.

The study was undertaken, utilizing Monte Carlo simulation of data to compare the three variable selection techniques in terms of the five criteria stated above. Fifteen populations were specified with predictor intercorrelations

operationally defined as low, medium, and high conjoint with a variety of predictor-criterion intercorrelation levels. For each population, using ten predictor variables and two cost functions, two hundred samples of N = 25, N = 50, and N = 100 were each drawn at random from the populations. Subsets of predictors were selected, using the sample data based on the three selection techniques. By computing $\hat{B}$ values for each technique, the $\bar{R}^2$, M.S.E., and the cost were calculated. Expected values of these criteria were computed for the many samples. The $\hat{B}$'s were then applied to the population to yield the average weight validity and the average mean squared error validity.

Results indicated that 1) contrary to expectations, the Bayesian and Ridge procedures yielded higher average $\bar{R}^2$ than the Forward Selection procedure for every sample size and for every intercorrelation level, although as sample size increased, the proportionate differences between procedures decreased; for example, at sample size = 25 the Ridge & Bayesian procedures yielded, on the average, an average $\bar{R}^2$ 15% higher than that of the Forward procedure but at sample size = 100, the difference decreased to 2.5%. 2) the Ridge Selection procedure yielded smaller mean squared error when the intercorrelation among predictors was high and/or as the sample size increased; for example, at sample size = 100, the Ridge procedure yielded, on the average, an average M.S.E. 27% smaller than that of the Forward procedure. 3) the Bayesian procedures, without exception, yielded subsets

having lower average costs than the other procedures; for example, the average cost 1 ($R.S.S._I + \frac{P}{k}\sqrt{R.S.S.}$) was, on the average, 37% less when the Bayesian 1 procedure was used rather than when the Forward procedure was used at sample size = 25. The average cost 2 ($R.S.S._I + \frac{P}{3k} R.S.S.$) was, on the average 69% less when the Bayesian 2 procedure was used in comparison with the Forward procedure at sample size = 25. 4) the Ridge procedure yielded subsets having a higher weight validity than the other procedures for all sample sizes and for all intercorrelation levels. Also, even the Bayesian procedures tended to yield higher average weight validity than the Forward Selection procedure; for example, at sample size = 50 the Ridge procedure yielded an I_1 9% larger than that of the Forward procedure and 1% larger than that of the Bayesian procedures. 5) when the sample size was large and/or the intercorrelation among predictors was high, the Ridge procedure produced subsets having a higher mean squared error validity than the other procedures. For example, at sample size = 100, the Ridge procedure yielded, on the average, an I_2 2% larger than that of the Forward procedure, 1% larger than that of the Bayesian 1 procedure, and 89% larger than that of the Bayesian 2 procedure.

Results were discussed and implications of the findings clearly indicated the useful properties of the Ridge and Bayesian Selection procedures for the researcher, as well as for the practitioner. Suggestions for future research were also offered.

Acknowledgements

I wish to express my deepest gratitude to the following individuals:

Dr. Alan Gross, Chairman of the committee with whom I conferred weekly and who guided me in the organizational and conceptual structure of this dissertation.

Dr. Damodar Gujarati, who always made himself available to me and who aided greatly in the theoretical understanding of the area.

Dr. Max Weiner, who made me aware of the practical meaningfulness of the dissertation.

Dr. R. Vinod, who served as a reader.

Dr. Donald Rock, a graduate professor of mine who first suggested the topic of this study to me and who always brought a sense of humor to technical or complex domains of educational psychological research.

Takahashi Hajime, who aided in the computer programming aspects.

Family and friends who shared my moments of joy and despair.

Table of Contents

List of Tables and Figures

List of Tables and Figures (Cont.)

List of Tables and Figures (Cont.)

Introduction

The educational and psychological researcher is often faced with the task of the selection of a subset of predictors from a larger set of predictors, in attempting to predict a criterion. For example, in order to predict college performance, the investigator may consider collecting information obtained from standardized test scores, high school grades, and biographical data. Several reasons exist why he may only wish to select a subset of variables from the numerous array of possible predictors. First, a reduction in the number of predictors facilitates computation. Second, reduction of predictors simultaneously reduces administration of tests used (where tests are the predictors) to predict the criterion. Third, this reduction of administrative time concomitantly reduces the overall monetary "cost" of testing. Fourth, and most importantly, reduction in the number of predictors will reduce the "problems" associated with applying regression coefficients derived from one sample in predicting the criterion variable in future samples.

At present there are numerous procedures which have been proposed to select a subset of predictor variables. However, their operating characteristics have not been fully investigated. Although each of the procedures was developed to satisfy some criterion for "good" variable selection, none of these procedures have been compared in terms of the other criteria. More specifically, it has not been shown how well

the various procedures work with respect to keeping down "cost," minimizing the residual sum of squares, minimizing the mean squared difference between population coefficients and sample coefficients, and also reducing the "problems" associated with prediction of the criterion in future samples.

It is the purpose of this research project to evaluate three procedures, Forward Selection, Ridge Regression, and a Bayesian Approach to Selection, with respect to the criteria of minimizing the residual sum of squares, minimizing the mean squared difference between population coefficients and sample coefficients, minimizing cost and reducing future sample prediction problems.

The Forward Selection procedure has been chosen for investigation because it is representative of Least-Square Estimation Procedures, i.e., variable selection procedures which make use of the traditional sample regression weights. These procedures minimize the residual sum of squares and are the most commonly used selection procedures among educational and psychological researchers. The Ridge Regression and Bayesian techniques are more innovative and have not been duly investigated.

The Ridge Regression procedure attempts to deal with correlations among predictors in a sample. It does so by attempting to diagonalize the intercorrelation matrix $X'X$ by augmenting the matrix by a constant k, thus making variable selection easier. The ridge estimates are biased estimators but have smaller expected mean square error than

traditional unbiased least-square coefficient estimates. The expected mean square error of an estimate means how close, on the average, is the squared distance of the estimate from the parameter which is being estimated.

The Bayesian selection procedure has been chosen because it uses a formal decision-theory approach that leads to a minimization of the cost factors. This procedure makes use of a conditional predictive distribution in attempting to predict future criterion scores from a subset of predictors. The investigator then considers the loss incurred when the actual criterion scores differ from the predicted criterion scores and the loss incurred when using a chosen set of predictors. Based on these losses, a decision may be made as to which predictor variables should be selected.

Therefore, each procedure will be compared in terms of the criterion which it uses to select variables and in terms of the two criteria which the other procedures use. In addition, all three procedures will be compared in terms of two other criteria, namely, reducing future sample prediction problems.

The results of this research will be useful to practitioners since they will provide guidelines for deciding which technique is most helpful to their specific prediction problems. For example, if the monetary cost is very important to the investigator, he may decide that if the Bayesian selection procedure minimizes this cost it is better than the other procedures. On the other hand, if an administrator

is concerned about the prediction of the criterion for future subjects, he may decide that the procedure which most consistently predicts the criterion is the best procedure to use.

Let us now consider in detail the notion of the cost and the notion of future sample prediction problems. Following this selection we will describe the logic of the three variable selection procedures.

Cost

The cost that is incurred when one chooses a set of predictors and then "combines" them to predict a dependent variable Y may be defined both as a non-psychometric cost and as a psychometric cost. The former is usually the cost incurred in collecting data on predictors. For example, in some cases such as predicting college performance, collecting test score data and interviewing prospective candidates incur differential costs--interviewing is far more costly than obtaining test score data. In other cases, the cost may be uniform if, for example, the predictors are each test item responses.

Psychometric cost refers to the subjective cost to the investigator of how important errors in prediction are to him. For example, suppose that actual college performance is Y = 10 (where 10 is an A college grade-point index), whereas the predicted performance is Y = 6 (where 6 is a C index). This difference is important. The importance is

the psychometric cost. We will use the sum of the squared difference between the actual and predicted scores as a measure of this psychometric cost (this will be discussed later).

The non-psychometric cost and the psychometric cost are combined to form the overall cost.

Problems Associated with Prediction in Future Samples

The basic problem associated with prediction in future samples may be demonstrated by a simple example. Suppose we are interested in developing a regression equation to predict a Y variable from the set of x variables (x_1, x_2,... x_p). We draw a random sample of size n and compute the sample regression equation $\hat{Y} = \hat{B}_o + \ldots \hat{B}_p x_p$, the squared correlation between the actual and predicted Y's $R^2(y,\hat{y})$, (i.e. the multiple correlation), and the residual sample variance $s^2_e = s^2_y(1-R^2)$. $R^2(y,\hat{y})$ and s^2_e are measures of how "accurately" the regression equation derived in our sample of size n predicts Y in this sample of size n. More specifically, $R^2(y,\hat{y})$ measures the variance shared by the actual and predicted Y's, and s^2_e measures the error in variance as a function of the correlation. However, what we are really interested in is the prediction of Y for future subjects, using the regression equation estimates found in the original sample. In other words, we wish to determine how "accurately" $\hat{Y}$ predicts Y for future subjects.

Browne (1969) has proposed two measures of the

predictive precision of a sample regression equation, in predicting future Y scores. The first measure is the correlation between the Y and $\hat{Y}$, where this correlation is taken over all future subjects. This correlation, which is conditional on the set of weights $\hat{B}$ computed in the original sample, can be expressed as:

$$(1A) \quad W^2(Y,\underline{X}|\hat{\underline{B}}) = \text{corr}^2\,(Y,\ \hat{Y}(\underline{X},\hat{\underline{B}})\ |\ \hat{\underline{B}}) = \frac{(\hat{\underline{B}}/\underline{\sigma}_{xy})^2}{\sigma_y^2 \cdot \hat{\underline{B}}/\underline{\Sigma}_{xx}\hat{\underline{B}}}$$

where $w^2(Y,\underline{X}|\hat{\underline{B}})$ is the squared correlation in the population future subjects of Y and $\hat{Y}$, given $\hat{\underline{B}}$, coefficient weights. That is, using coefficients obtained in the original sample, w^2 is the squared correlation between the predicted Y's and the Y's of the population of future subjects.

where $\hat{\underline{B}}$ is the vector of coefficient weights obtained from the original sample,

where $\hat{Y}(X|\hat{\underline{B}})$ is the predicted $\hat{Y}$ in the population, given $\hat{\underline{B}}$,

where $\underline{\sigma}_{xy}$ is the covariance of x and Y in the population,

where σ_y^2 is the variance in the population, and

where $\underline{\Sigma}_{xx}$ is the X/X correlation matrix in the population.

That is, the squared cross-validity coefficient, or weight validity, is the correlation between Y scores in the population future subjects and the predicted Y scores $(\hat{Y})$. The $\hat{Y}$ scores are estimated using the predictors and their respective coefficients obtained in the original sample.

The second measure of predictive precision of a sample regression equation for future subjects is the mean squared error in a sample, given coefficients estimates $\hat{B}$ computed in the original sample, expressed as:

$$(1B) \quad D^2(Y, \underline{X}|\hat{\underline{B}}) = E\,([Y-\hat{Y}(\underline{X}|\hat{\underline{B}}]^2|\hat{\underline{B}}) = \sigma^2(Y|\underline{X}) + (\hat{\underline{B}}-\underline{B})'\underline{\mu}_x + (\hat{\underline{B}}-\underline{B})\underline{\Sigma}_{xx}(\hat{\underline{B}}-\underline{B})$$

where $\sigma^2(Y|\underline{X})$ is the conditional variance of Y, given X scores in the population,

where $(\hat{\underline{B}}-\underline{B})$ is the difference between predicted and actual $\underline{B}$ in the population. The predicted $\hat{\underline{B}}$ were obtained from the original sample,

where $\underline{\mu}_x$ is the vector of x means in the population, and

where $\underline{\Sigma}_{xx}$ is the covariance matrix of x variables.

That is, the mean squared error is equal to the expected squared difference between the actual Y scores and the predicted Y scores in the sample, using coefficients obtained in the calibration sample.

One may consider taking an infinite number of samples

to determine the expected value of W^2 and D^2. The population multiple correlation, $\rho^2 = corr^2(Y,Y_{pop})$, sets an upper bound to $E(W^2(\underline{Y},\underline{X}))$. In other words, if our original sample size was ∞ and $\hat{\underline{B}} = \underline{B}$, then we would achieve this maximum prediction process in a correlational sense. Similarly, the lower bound to $E(D^2(Y,X))$ is the mean squared error achieved in the population, using its population regression equation, i.e. $\sigma^2_{y|x} = E(Y-B_o-B_1x_1....-B_px_p)^2 = \sigma^2_y(1-\rho^2)$. By relating $E(W^2)$ with ρ^2 and $E(D^2)$ with $E(Y-Y_{pop})^2$ we are scaling the quantities.

Therefore, it is proposed that two indices be used to evaluate the three selection techniques.

$$(2A) \quad I_1 = \frac{E(W^2)}{\rho^2} \leq 1$$

This is the expected value of the weight validities obtained in (1A), divided by the population multiple correlation. Index I is bounded above by 1, and the closer it is to one, the more precise the prediction.

$$(2B) \quad I_2 = \frac{E(Y-\hat{Y}_{pop})^2}{E(D^2)} = \frac{\sigma^2_y(1-\rho^2)}{E(D^2)} \leq 1$$

This is the population mean squared error, divided by the expected value of the mean squared errors obtained in (2A). Index I_2 is bounded above by 1 (since $E(D^2)$ is always greater than the population mean squared error), and the closer it is to one, the more precise the prediction.

The indices I_1 and I_2 will be used to evaluate the

future sample predictive precision in the Forward Selection, Ridge Regression, and Bayesian selection of variables. These indices will be used in addition to the criteria of cost, residual sum of squares, and the mean squared difference between population coefficients and sample coefficients.

Forward Selection

In the Forward Selection procedure, where $x_1, x_2, \ldots x_p$ variables are used to predict Y, one first considers the correlation between Y and each x_i (r_{yxi}), for i = 1, 2, ... p. One then chooses the variable (xj) which has the highest correlation with Y and asks whether this variable adds significantly to its regression equation. In other words, in the model $Y = B_o + B_j x_j + e$, is $B_j = 0$?

In order to test for significance, an F-test is applied, setting an F value as the criterion. The form of the F-test is as follows:

$$(3) \quad F = \frac{r^2_{yxj}}{(1-r^2_{yxj})/(n-2)}$$

where n = number of subjects.

If the difference is significant, variable j enters the model and one then considers whether a second variable can enter the model. We examine the semi-partial correlations

of the remaining variables with Y, with variable X_j partailed out. The variable with the highest semi-partial correlation (e.g., x_k) is added to the model and this other model is compared with the model excluding that variable. More specifically, in the model $Y = B_o + B_j x_j + B_k x_k + e$ we test $H_o : B_k = 0$.

Again an F-test is applied to test for significance. The F-test is of the form:

$$(4) \quad F = \frac{r^2_{yx_{jk}} - r^2_{yxj}}{(1-r^2_{yx_{jk}})/(n-2)}$$

where $r^2_{yx_{jk}}$ is the correlation of x_j and x_k with Y, and

where r^2_{yxj} is the correlation of x_j with Y.

This processs is continued until a non-significant F is encountered.

In essence, this procedure finds the smallest number of predictors such that the residual sum of squares will not be significantly increased.

To summarize, if the variance reduction obtained by adding the most highly partially correlated variable to a regression equation is significant at a specified F level, the variable is included in the model. If the variance reduction of the most highly partially correlated variable in a regression is insignificant at a specified F level, this variable is not added to the regression and the process terminates. The F-test, used to test for significance, is

of the form:

$$(5) \quad F = \frac{r^2_{y\cdot 12.....p} - r^2_{y.....k}}{(1-r^2_{y.12.....p;})/n-2}$$

where 12.....p are the variables being tested in the new model,

where 12....k are the variables found to be significant in the previous model, (as in equation 4) and

where n is the sample size.

Ridge Regression Selection Procedure

Ridge Regression estimation and selection procedure was first developed by Hoerl and Kennard (1970) and may be viewed as a modification of least-squares estimation procedure. The least-squares estimation procedure will be reviewed prior to a discussion of Ridge estimation.

Least-Squares Estimation Procedure

In the least-squares procedure, one considers p predictors used in a sample to predict Y in the regression model:

$$(6) \quad \underline{Y} = \underline{XB} + e$$

where $\underline{Y}$ is an (nxl) vector of the criterion,

where $\underline{X}$ is an (nxp) vector of predictor

variables.

where $\underline{B}$ is a (pxl) vector of population regression coefficients, and

where $\underline{e}$ is a (nxl) vector of residual errors.

The least-square estimates minimize the sum of the squared errors (SS_e) in the sample (otherwise known as the residual sum of squares). The sum of squares is represented as:

$$SS_e = \Sigma (Y-\hat{B}_o-\hat{B}_1x_1-\ldots.-\hat{B}_px_p)^2$$

or, alternately, in matrix form, as:

$$(7) \quad \emptyset(B) = (\underline{Y}-\underline{X}\underline{\hat{B}})' (\underline{Y}-\underline{X}\underline{\hat{B}})$$

The solution for $\hat{B}$ is:

$$(8) \quad \underline{\hat{B}} = (X'X)^{-1}X'Y$$

where $\underline{\hat{B}}$ is the (pxl) vector of coefficient estimates.

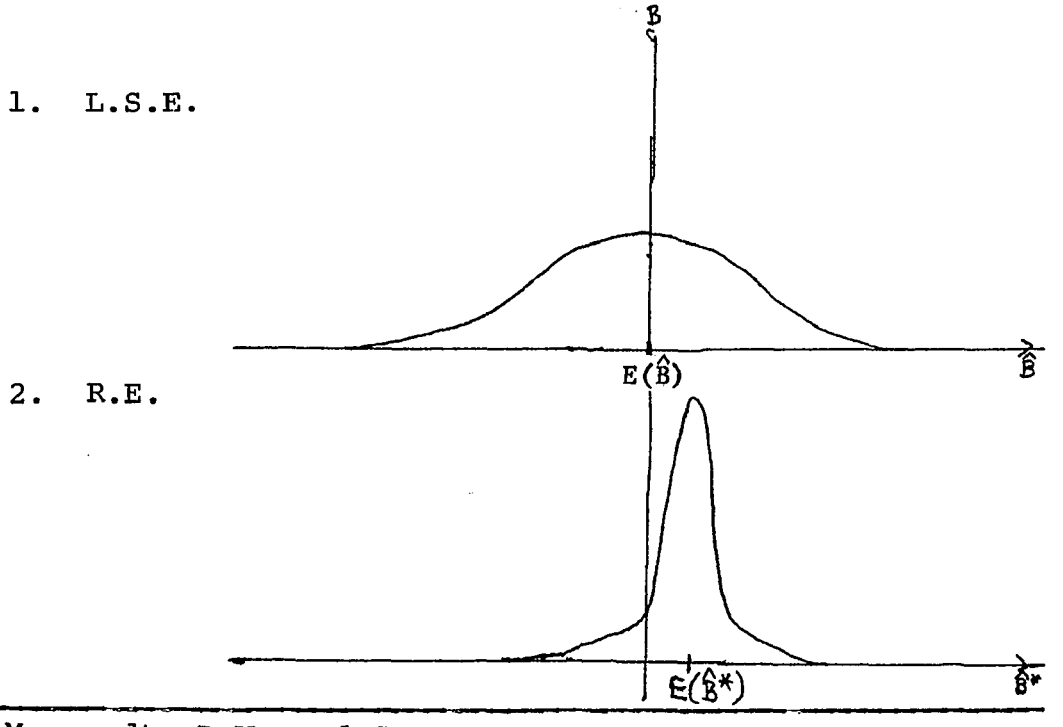
The properties of these estimates are that they are unbiased, i.e. $E(\underline{\hat{B}}) = \underline{B}$, and that, in comparison to all other linear unbiased estimates of B, the least-squares (L.S.) weights have minimum sampling variance. However, it is important to note that there exist estimates which, although

biased, are superior in some sense to the L.S. estimates. In terms of these sampling variance concepts, Ridge regression in fact provides such estimates.

Ridge Estimation

Ridge estimation can be viewed as accomplishing two purposes. First, it provides estimates $(\hat{B}_i^*)$ of the coefficients (B) that have a "higher probability" of being closer to the coefficients (B) than L.S. weights. To better understand this concept, let us consider the expected squared distance $E(\Sigma(B_i-\hat{B}_i)^2) = E(L^2)$ for both L.S.E. and Ridge estimates. The following displays the sampling variance and bias of a least-square estimate $(\hat{B}_i)$ and the variance and bias of a Ridge estimate $(\hat{B}_i^*)$:

Figure 1: Comparison of L.S.E. and R.E.
With Respect to the Bias and Variance



Source: Marquadt, D.W. and Snee, R.D., 1975, p. 5.

In the L.S.E., $E(\hat{B}) = B$, i.e. the expected value is equal to the population parameter. Hence $\hat{B}$ has zero bias. However, it may be observed that the variance is large, which means that the probability of $\hat{B}_i$ being far from B_i is quite high. On the other hand, in the Ridge estimate, $E(\hat{B}^*) \neq B$, so there is bias in the estimate. But it can also be observed that the variance is small, so the probability that $\hat{B}_i^*$ is far from B_i is small indeed. It will be shown that $E(L^2)$, the mean square error of an estimate, is equal to the sampling variance plus the squared bias of the estimates. Hence, if one is willing to accept some bias, one may choose estimates, other than L.S.E., such as R.E., which have a smaller variance and which are closer to B. This notion of closeness is quantified in terms of the concept of expected squared distance $E(L^2)$, synonomous with the mean squared error (MSE).

A second goal accomplished by Ridge Regression is a reduction of intercorrelations among predictors, making the prediction system behave more like an orthogonal system. In this manner it makes it clear how variables can be selected by simply looking at the coefficients ($\underline{\hat{B}}^*$) to choose the best variables. Let us now consider a formal comparison of L.S.E. with R.E. to demonstrate the general instability (large sampling variance) associated with L.S.E. in comparison to R.E.

The variance-covariance matrix in a sample of the joint distribution of $\hat{B}$ is given as:

$$(9) \quad \mathrm{var}(\hat{\underline{B}}) = \sigma^2 (X'X)^{-1}$$

where σ^2 is the population variance, and where $X'X$ is the standardized raw score matrix so that it is now an intercorrelation matrix.

Let L = the distance from $\hat{B}$ to B in a sample. The square of L is $L^2 = (\hat{\underline{B}}-\underline{B})'(\hat{\underline{B}}-\underline{B})$.

The expectation over samples is:

$$(10) \quad E(L^2) = \sigma^2 \mathrm{Trace}(X'X)^{-1}$$

Since it can be shown that

$$(11) \quad E(\hat{\underline{B}}'\hat{\underline{B}}) = \underline{B}'\underline{B} + \sigma^2 \mathrm{Trace}(X'X)^{-1},$$

we can express $E(L^2)$ as:

$$(12) \quad E(L^2) = E(\hat{\underline{B}}'\hat{\underline{B}}) - \underline{B}'\underline{B}$$

As $E(L^2)$ becomes smaller, the expectation of the squared regression vector $E(\hat{\underline{B}}'\hat{\underline{B}})$ approaches $\underline{B}'\underline{B}$.

For the error $\underline{e}$ distributed multivariate normal with zero means, the variance of L^2 is:

$$(13) \quad \mathrm{var}(L^2) = 2\sigma^4 \mathrm{Trace}(X'X)^{-2}$$

The value of $E(L^2)$ can be expressed in terms of the eigenvalues of $X'X$. These eigenvalues are:

$$(14) \quad \lambda_{max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p = \lambda_{min} > 0$$

The above signifies that the eigenvalues, or characteristic roots of the $X'X$ matrix, may be ordered from highest to lowest, where λ_{max} is the largest eigenvalue and where λ_{min} is the smallest eigenvalue.

The average squared distance $E(L^2)$ from $\hat{B}$ to B, in terms of these eigenvalues is:

$$E(L^2) = \sigma^2 \Sigma (1/\lambda_i).$$

When the $X'X$ matrix has one or more small eigenvalues, which often occurs for highly correlated predictors, the distance from $\hat{B}$ to B will tend to be large in expectation and will also tend to have large variance; i.e., $E(L^2)$ will be large.

This is true due to the lower bounds of $E(L^2)$ and $var(L^2)$, which are σ^2/λ_{min} and $2\sigma^4/\lambda^2_{min}$, respectively. When the matrix $X'X$ has one or more small eigenvalues, the distance from $\hat{B}$ to B will be large. Whenever the predictors are correlated, there is a greater probability that small eigenvalues would be obtained.

Marquadt and Snee (1975) define the variance inflation factor for each term in a regression model as a measure of the collective impact of the intercorrelations on the variance of the coefficient of that term. These variance inflation factors are the diagonal elements of the inverse of the simple correlation matrix. The authors empirically demonstrate how least-square estimators, in an attempt to yield unbiased estimators, often produce coefficients with

large variance inflation factors which do not predict well in new data.

To summarize, the average squared distance from $\hat{B}$ to B may be large when using L.S.E. for highly correlated predictors.

Derivation of Ridge Estimators

The Ridge Estimation procedure augments the main diagonal elements of the intercorrelation matrix $X'X$ by a constant to yield $[X'X + KI]$. The derivation of this technique can be found in Appendix C.

Let us now compare the characteristics of the Ridge estimators with those of the L.S. estimators.

Comparison of the Characteristics of Ridge Estimators With Least-Square Estimators

When the x's are highly correlated, i.e. when $X'X$ has one or more small eigenvalues, the expected distance $E(L^2)$ will be large since the sampling variance of each $\hat{B}$ will be large in this case. However, by substituting $X'X + kI$ for $X'X$, where k is a real number ranging from 0 to ∞, stabilization of coefficients in terms of smaller numerical values may be achieved. Thus $E(L^2)$ for R.E. (Ridge Estimators) will be less than that for L.S.E. This relationship between $E(L^2)$ for L.S.E. and $E(L^2)$ for R.E. (herein termed $E(L^{*2})$) can be found in Appendix D.

Another difference between the R.E. and the L.S.E. is that the R.S.S. (residual sum of squares) of the L.S. estimates in the sample is $\emptyset = (\underline{Y}-\underline{X}\hat{\underline{B}})'(\underline{Y}-\underline{X}\hat{\underline{B}})$, whereas for the R.E. it is $\emptyset^*(k) = (\underline{Y}-\underline{X}\hat{\underline{B}}^*)'(\underline{Y}-\underline{X}\hat{\underline{B}}^*) = (\underline{Y}-\underline{X}\hat{\underline{B}})'(\underline{Y}-\underline{X}\hat{\underline{B}}) + (\hat{\underline{B}}^*-\hat{\underline{B}})'X'X(\hat{\underline{B}}^*-\hat{\underline{B}})$, where the first term on the right hand side of the equation is the L.S.E. R.S.S. and the second term is the increment in R.S.S. due to using the R.E. Therefore the R.S.S. of the Ridge estimates is always equal to or greater than the R.S.S. of the L.S.E.

To reiterate, it has been proven that there exists $k > 0$, which yields a mean square error ($E(L^2(k))$ which is smaller for R.E. than that for the L.S.E. In other words, by choosing $k > 0$ one can produce estimators (R.E.) which have the property that their mean square error is smaller than that of L.S.E.

By examining the mean square error properties of the Ridge estimator $E(L^2(k))$, one can qualitatively interpret the relationship between the variances of the parameter estimates, the squared bias, and the parameter k. The mean square error of the Ridge estimates is:

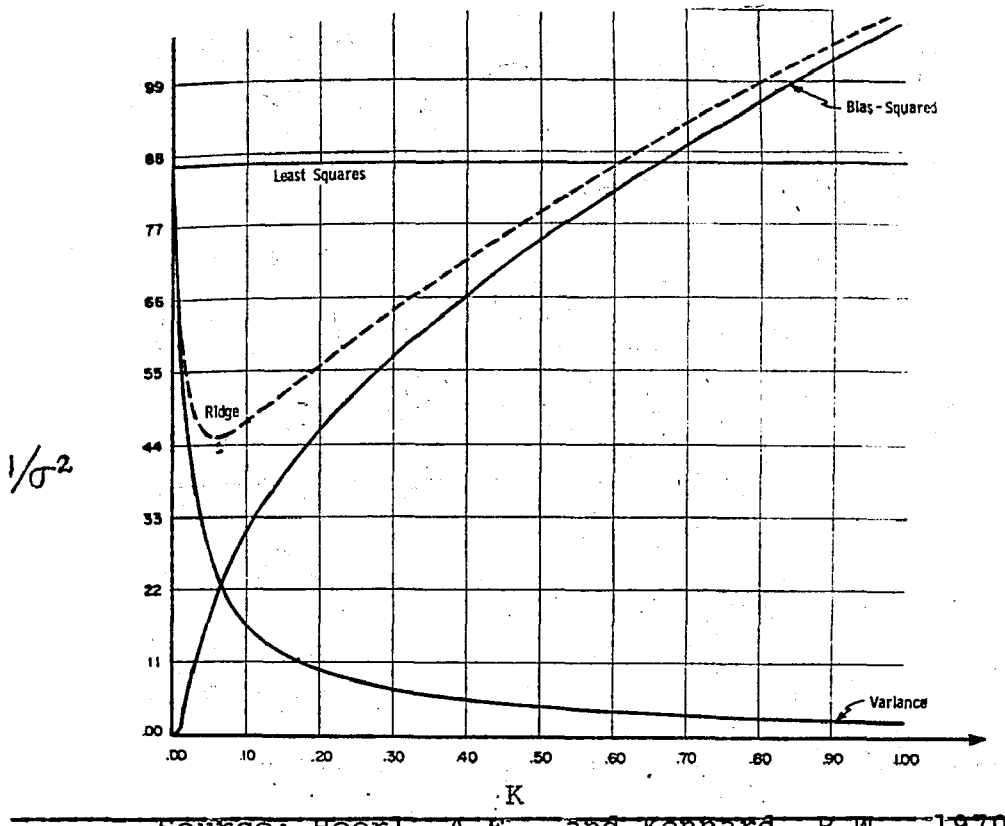
$$(15) \quad E(L^{2*}) = E[L^2(k)] = E[(\hat{\underline{B}}^*-\underline{B})'(\hat{\underline{B}}^*-\underline{B})]$$

The mean square error is the sum of the variances of the parameter estimates and the squared bias. The variance decreases monotonically as k increases, while the bias increases monotonically as k increases. Hence in choosing

$k > 0$, one may reduce the variance, despite some bias, thereby minimizing the mean square error.

To illustrate this, one may plot mean square error functions for different values of k to compare Ridge estimator mean square error with L.S.E. mean square error, as follows:

Figure 2: Mean Square Error Functions



Source: Hoerl, A.E., and Kennard, R.W., 1970, p.61.

The horizontal line is the mean square error of the L.S.E. The broken line is the mean square error of the R.E. The monotonically increasing curve is the squared bias which can be observed to increase as k increases. The monotonically decreasing curve is the variance which decreases as k increases. The mean s.e. of the L.S.E. is a constant

since it does not change for different values of k. On the other hand, increasing k slightly minimizes the mean square error of the R.E. since the mean square error is the sum of the variance and the bias. It may also be observed that there are many values of k for which the mean square error of the R.E. is less than that of the L.S.E.

Choosing a Value of K

While the value of k cannot be determined explicitly from the theoretical formulation of the mean squared error function, more than one strategy has been proposed to select a value for k.

Hoerl and Kennard (in written communications, 1975) propose estimating k as:

$$(16) \quad k = \frac{p\hat{\sigma}^2}{\underline{\hat{B}}'\underline{\hat{B}}}$$

Using a computer-simulation of data, with the estimate of k cited, they report that when using the above estimate the mean square error is uniformly less for Ridge estimators than for least-square estimators, under a variety of data-generating parameters.

An alternate procedure suggested by Hoerl and Kennard (1970) for estimating k is to examine the Ridge trace, a two-dimensional plot of the $\hat{B}_i^*(k)$ and the residual sum of squares, $\emptyset^*(k)$, for different values of k, usually within

the interval of 0 to 1. The Ridge trace displays the R.E. coefficients for different values of k. A value of k is chosen which is the smallest value for which coefficient estimates do not change drastically, thereby reducing the variance of the estimates without too much bias.

In summary, the following practical benefits from Ridge estimation procedures may be accrued:

1) For a particular k value, the system will stabilize (i.e. the variances of the parameter estimates will decrease) and have characteristics of orthogonal systems such as non-small eigenvalues. That is, although the predictors are highly correlated in the sample, choosing $k > 0$ reduces the correlation between predictors so that the system behaves like an orthogonal system. This allows the researcher to observe which predictors are best included in the prediction system.
2) Regression coefficients will not have unreasonable absolute values with respect to variables which are highly intercorrelated. That is, by augmenting the $X'X$ matrix, the squared regression coefficient vector $\hat{\underline{B}}^{*\prime}\hat{\underline{B}}^{*}$ will be smaller.
3) Coefficients with incorrect signs when $k = 0$ may change to the proper sign. Often in real data problems, coefficients are observed with a negative sign which the researcher knows should

be positive. The R.E. procedure corrects for this.

4) The R.S.S. may not inflate to unreasonable value, i.e. it may not be large relative to the minimum R.S.S. By comparing the Ridge trace at k = 0 with other values of k, the researcher is cautious not to choose too large a k, for this would have a large R.S.S. and hence too much bias.

Ridge Regression: An Example

Two empirical studies analyzed by Hoerl and Kennard (1970) demonstrate how the Ridge trace can be utilized for a given set of data. The first study, utilizing Gorman and Toman's (1966) data, reveals that $E(L^2)$ = more than three times what would have been obtained for an orthogonal system. That is, if the $X'X$ matrix were uncorrelated and hence a matrix with O's in the off-diagonal, $E(L^2)$ would be $p\sigma^2$. The second study, an analysis of Jeffers (1967) data, also leads Hoerl and Kennard to conclusions identical to that of the first study.

For the Gorman and Toman data (see Appendix A), Hoerl and Kennard compute the eigenvalues of $X'X$ as follows:

$\lambda_1 = 3.692$	$\lambda_5 = .972$	
$\lambda_2 = 1.542$	$\lambda_6 = .659$	$\lambda_9 = .152$
$\lambda_3 = 1.293$	$\lambda_7 = .357$	$\lambda_{10} = .068$
$\lambda_4 = 1.046$	$\lambda_8 = .220$	

$$\sum_{i=1}^{p}(1/\lambda_i) = 33.825$$

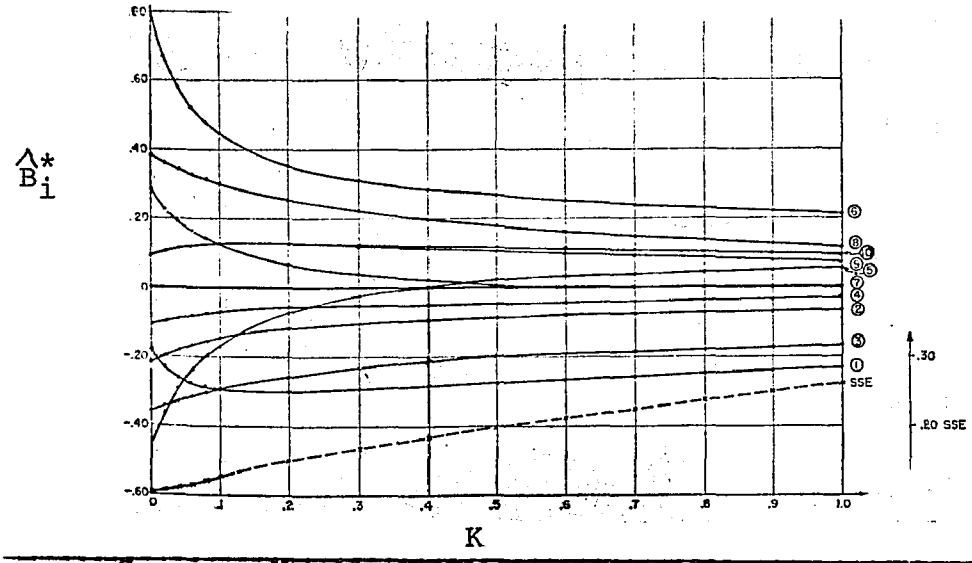
Therefore, $E(L^2) = 33.825\sigma^2$

This is more than three times what it would be for an orthogonal system since $\sum_{i=1}^{p}(1/\lambda_i) = p$ for an orthogonal system and $E(L^2) = 10\sigma^2$. This is due to a large number of significant interfactor (factor is used synonomously with variable) correlations which are reflected in the eigenvalues.

The Ridge trace is as follows:

Figure 3: Ridge Trace With Ten Variables

(circled numbers refer to the respective variables)



Source: Hoerl, A.E., and Kennard, R.W., 1970, p. 71.

The different values of $\hat{B}^*$ displayed in the Ridge trace have been computed using different values of k for $\hat{B}^* = [X'X + kI]^{-1} X'Y$.

At k = 0, the L.S.E. coefficients appear to be over-estimated since they are larger than for $k > 0$; also, since

the estimates change considerably for different values of k, they can be said to be unstable. Variable 5 changes sign for $k > 0$, and variable 6 decreases very rapidly; variable 1 seems to have been underestimated at $k = 0$ since it increases for $k > 0$. Variable 7 also seems to be overestimated and approaches zero. Variable 5 has a negative coefficient with the largest absolute value. But for $k > 0$, variable 5 goes to zero and eventually becomes positive. Variable 6 also decreases rapidly but stabilizes and does not go to zero. Variables 5 and 6 have a correlation of .84, which indicates that they are almost identical. Thus it is surprising that their effects are opposite in sign. This is due to a covariance of -4.33 which forces them apart to produce opposite signs. At $k = 0$, variable 1 is the second least important negative factor. But as $k > 0$, it increases in absolute value. This is due to the correlation of variable 1 with the other variables. Since the other negative variables are overestimated at $k = 0$, for $k > 0$ variable 1 becomes the most important negative variable. Variable 7 is overestimated and goes to zero.

Thus the Ridge trace portrays, as the intercorrelation among the x's is decreased, i.e. k is increased, which variables do not hold their predictive power. One can observe which variables change drastically or whose effect is negligible in the overall regression equation and are not likely to be effective predictors in other samples.

For $.2 < k < .3$, the prediction system is stable and

may yield estimates closer to B than L.S.E. coefficients. Therefore, by choosing k = .25, one obtains the following Ridge estimates for the unknown population weights ($\underline{B}$):

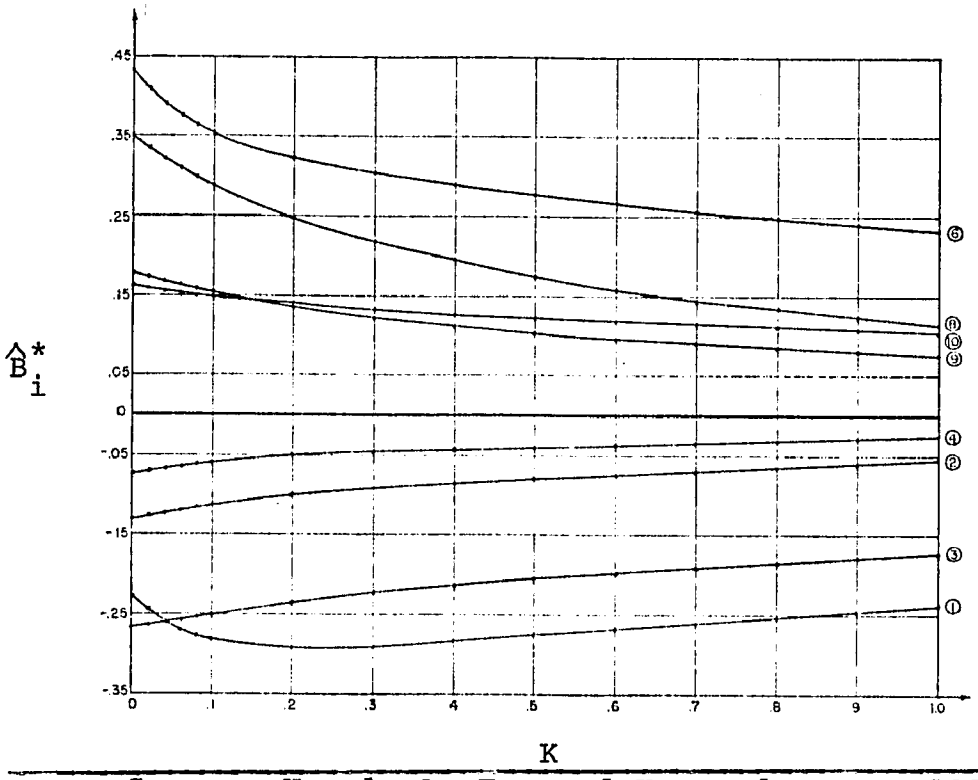
$$\hat{B}_1^* = -.295 \qquad \hat{B}_6^* = .325$$
$$\hat{B}_2^* = -.110 \qquad \hat{B}_7^* = .05$$
$$\hat{B}_3^* = -.245 \qquad \hat{B}_8^* = .24$$
$$\hat{B}_4^* = -.05 \qquad \hat{B}_9^* = .125$$
$$\hat{B}_5^* = -.04 \qquad \hat{B}_{10}^* = .125$$

Ridge Regression Subset Selection: An Example

In order to determine which factors should be deleted, one examines which variables or factors do not hold their predictive power, i.e. are not stable across k. While several criteria (see p. 26) exist for determining which variables are not stable, Hoerl and Kennard use one criterion, as follows: Factors 5 and 7 are unstable since, for $k > 0$, they approach zero. A new Ridge trace, with factors 5 and 7 deleted, is:

Figure 4: Ridge Trace, Ten Factor Example With Variables 5 and 7 Deleted

(circled numbers refer to variables)



Source: Hoerl, A. E., and Kennard, R.W., 1970, p. 74.

Overestimations and instabilities present in the full system are dampened when these variables are deleted. Hence the new subset would not include variables 5 and 7.

Three criteria may be applied to discard variables:

1) Those variables that converge near zero (since they are ineffective),
2) Those variables that change signs (since they are unstable), and

3) Those variables that drastically change in absolute value (since they, too, may be unstable).

Any or all of the above criteria may be used to delete variables. Once the variable has been discarded, the investigator may either a) choose k = 0 in estimating the coefficients such that $\hat{\underline{B}}^* = (X'X)^{-1}X'Y$, where $X'X$ is the reduced correlation matrix, after the variables have been deleted, or b) refine his estimation by computing a new Ridge trace to find a value of k where the coefficients are stable, or c) refine his estimates by using an explicit estimate of k such as $\frac{p\hat{\sigma}^2}{\underline{B}^{*\prime}\underline{B}^*}$ to augment the reduced $X'X$ matrix (i.e. the original matrix which, let us say, consisted of ten predictors, now consists of six predictors).

We will use procedure (c) as our Ridge estimation procedure. This procedure first uses the Ridge trace to select variables and then uses an estimate of k to refine the coefficient estimates. This procedure has been chosen because the Ridge trace allows one to select variables, and once they have been selected, an estimation of k, to refine the variables, is easy to apply.

Bayesian Selection of Predictors

The third approach, a Bayesian one, was conceptualized by Lindly (1968), who defined the problem of selecting variables for the prediction of a criterion in a Bayesian framework. To provide the reader unfamiliar with the principles

of Bayesian statistics with a rudimentary understanding of this statistical theory, a brief description ensues.

Bayesian Statistics

Suppose one wished to determine the number of students in some population who require remediation. In attempting to answer this question, the classical statistician would draw a random sample from the population to determine the percentage which needs remediation and use that sample percentage as an estimate of the number of students in the population who require remediation. He could either use that percentage as a point estimate, place confidence intervals about that estimate, make hypotheses, or use some combination of these alternatives.

The Bayesian statistician would also represent the probability of observing his sample results in the form of a likelihood function. The likelihood function refers to the data-generating process or sample data. This distribution is the probability of obtaining these sample data, if the parameter were known. However, he could also assess his a priori beliefs about what the parameter values are in the form of some probability distribution. This prior distribution refers to the discrete or continuous probability distribution, or simply the beliefs about the parameter values, prior to current observations (sample data). This prior distribution may be subjective. Classical statisticians take issue with the use of subjective priors, which they

feel are not objective measures. Bayesians, on the other hand, rebut that the classical statisticians also implicitly assign prior values, while only Bayesians quantify these values. When a priori beliefs are uncertain, the prior distribution may be represented as a uniform distribution. Such a distribution is non-informative and has been termed "indifferent" or "diffuse" relative to sample data. In such a case, the sample data "swamps" the prior distribution because results are solely based on sample data. By combining a priori beliefs with sample data, the investigator computes a posterior probability distribution, which enables him to make probabilistic statements about the parameters, given the data. Although the results will not differ from the classical approach when a diffuse prior is used, interpretation of the results will differ.

In the example above, suppose that a sample of five was drawn in which one student needed remediation ($p = .20$). For simplicity, let us assume that the percentage (P) in the population which needs remediation can only be one of four values: .01, .05, .10, or .25. Table 1 indicates for different values of P, the prior probabilities of such values, the likelihood and the posterior probabilities.

Table 1: Bayesian Example of a Probability Distribution

P	Prior Probability	Likelihood	Prior x Likelihood	Posterior Probability
.01	.600	.0480	.0288	.232
.05	.300	.2036	.0611	.492
.10	.080	.3280	.0262	.212
.25	.020	.3955	.0079	.064
	1.000		.1240	1.000

For example, the investigator may feel that the probability of P = .01 is .60. Upon observing his data, the researcher finds that the probability of P = .01 is .0480. The conditional posterior probability of P = .01 is found to be .232. Thus the incorporation of sample data has altered his belief (from .0480 to .232) as to the probability that P = .01.

The investigator can now use these revised probabilities in making inferences or decisions (this will be discussed later). In addition to the posterior distribution, the investigator may also be interested in making predictions about a future sample outcome before this outcome is actually observed. These probabilities associated with future outcomes are expressed in the form of a probability distribution known as the predictive distribution. In the example above, suppose that one wished to determine the probability that an (n + 1)st, i.e. a sixth, subject would need remediation. This probability would be the sum of the

products of the likelihood that the subject would need remediation, given a state-of-the-world, and the posterior probabilities. Thus, the posterior predictive probability that the future outcome would need remediation is:

$$\begin{aligned}\text{Pred (future outcome = 1)} &= L(0=\text{outcome}=1\ P=.01)(P''(P=.01) \\ &\quad + L(0=1\ P=.05)P''(P=.05) + \\ &\quad L(0=1\ P=.10)P''(P=.10) + \\ &\quad L(0=1\ P=.25)P''(P=.25) \\ &= (.01 \times .232) + (.05 \times .492) + \\ &\quad (.10 \times .212) + (.25 \times .064) \\ &= .0850\end{aligned}$$

L = likelihood

The posterior predictive probability that a sixth subject selected would need remediation is therefore .085.

In suing the posterior or predictive distribution to make inferences, the Bayesian can employ a formal decision-theory framework. By combining knowledge concerning the values of the parameters (states-of-the-world) with the decisions to be made under respective circumstances, the Bayesians may express a utility function. This utility represents the intrinsic importance to the decision-making of any decision occurring in conjunction with any state-of-the-world.

In the example cited previously, the state-of-the-world is the percentage of students needing remediation. The actions may be to remediate the population or not to remediate. Suppose the utility of any decision is expressed in Table 2.

Table 2: Utility Function: An Example

		State-of-the-World (percentage needing remediation)			
		.01	.05	.10	.25
Action:	Remediate	0	1	10	100
	Do Not Remediate	100	10	1	0

The utility of remediating the population can be observed to increase as the percentage of students needing remediation increases. On the other hand, the utility of not remediating decreases as the percentage of students needing remediation increases. In order to determine the expected utility of any decision, one sums the products of the probability of a state-of-the-world with its respective utility. One can then choose the decision which has the highest expected utility. Since the utility of a decision may be defined as the converse of the loss of a decision, one seeks to choose the decision which has the lowest expected loss.

In conclusion, there exist salient differences in outlook between the Bayesian and non-Bayesian statistics. In the former, parameters are random variables, while in the latter, statistics are estimates of parameters of fixed value. The Bayesian may incorporate his prior knowledge about the parameters by combining prior distribution with current sample data to form posterior distributions of the parameters. When

his prior knowledge is diffuse, or indifferent relative to current sample data, the use of an indifferent prior in combination with sample data yields distributions similar to those obtained by non-Bayesians.

We shall now employ this Bayesian framework in discussing our specific problem, that of selecting a subset of variables.

Bayesian Selection

In the Bayesian approach to variable selection, one considers the prediction of the Y scores of a future sample of subjects from a set of X scores. The problem of variable selection is to decide which X variables we should observe and which X variables we should not observe in the future sample. Thus, the predictive distribution of the Y's and X's will play a central role in the Bayesian model. Let us now consider each step of the Bayesian model.

Initially we observe a sample from the population. It is assumed that the data-generating process, or likelihood function, is a multivariate normal distribution with parameters $\underline{\mu}$ and $\underline{\Sigma}$. Thus, k x-variables are weighted with coefficients $\underline{B}$ so as to predict the Y variables. Then the researcher assesses his prior beliefs in the form of a prior distribution. Let us assume that our prior beliefs are not strong about what the mean vector $\underline{\mu}$ of $\underline{X}$ and Y, as well as the variance-covariance matrix $\underline{\Sigma}$ of Y and $\underline{X}$, are. The prior distribution would be represented as a diffuse prior. The

likelihood function may then be combined with the prior distribution to yield a posterior distribution of $\underline{\mu}$ and Σ, given the mean $\underline{m}$ and variance-covariance V of the sample.

In order to determine the predictive distribution of future outcomes of $\underline{X}$ and Y scores, a joint probability of future outcomes ($\tilde{\underline{Y}}$ and $\tilde{\underline{X}}$) is established. The posterior predictive distribution of future outcomes is obtained by integrating over the values of $\underline{\mu}$ and Σ. This distribution is a multivariate t-distribution. The conditional distribution of Y, given X_1, X_2, ... X_k is also a t-distribution. Given the value for k potential predictor variables (X_1, ... X_k) for a future subject, a decision needs to be made as to how this X information will be used to predict Y. We can denote the predicted Y in general as $f(X_1, \ldots X_k)$. The loss resulting from a given decision consists of two parts. The first is the monetary loss (incurred by using a chosen set of predictors). The second component is a psychometric loss which results when the actual Y does not equal $f(X_1, \ldots X_k)$. A quadratic loss function $(Y-f(X_1, \ldots X_k))^2$ will be used to represent this loss.

The problem may now be formulated more precisely in the following manner: Since a priori strengths of beliefs about the parameters are vague, the prior probability distribution function (p.d.f.) of the mean vector ($\underline{\mu}$) is diffuse and may be represented as:

$$(17) \quad P(\underline{\mu}) \propto \text{constant}$$

The prior p.d.f. of the variance-covariance matrix is also

diffuse and is represented as:

$$(18) \quad P(\Sigma) \quad \frac{1}{|\Sigma|^{k/2}}$$

The joint prior p.d.f. of $\underline{\mu}$ and Σ is given as:

$$(19) \quad P'(\underline{\mu},\Sigma) \propto \frac{1}{|\Sigma|^{\frac{k+1}{2}}}$$

The likelihood, upon drawing a sample of X's and Y's from a multivariate normal distribution, may be represented as:

$$(20) \quad \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{(k+1)}{2}}|\Sigma|^{1/2}} \exp[-\tfrac{1}{2}(\underline{0}_i-\underline{\mu})'\Sigma^{-1}(\underline{0}_i-\underline{\mu})]$$

where $\underline{0}_i = \begin{bmatrix} Y_i \\ \underline{X}_i \end{bmatrix}$, the individual Y and $\underline{X}$ scores.

By combining the prior p.d.f. with the likelihood function, one finds that the joint posterior probability of $\underline{\mu}$ and Σ, $P''(\underline{\mu},\Sigma)$, is a Wishart distribution, the marginal distribution of $\underline{\mu}$ is a multivariate normal distribution, and the marginal distribution of Σ is a Wishart distribution.

It is necessary to determine the predictive distribution of a future set of n observations of Y and $\underline{X}$ from a multivariate normal distribution, given the results of the previous n observations.

This future data is denoted as:

$$\tilde{\underline{Y}} = [\tilde{Y}_1, \tilde{Y}_2, \ldots \tilde{Y}_{n*}]$$

$$\tilde{\underline{X}} = [\tilde{X}_1, \tilde{X}_2, \ldots \tilde{X}_{n*}]$$

The posterior predictive probability is the product of the posterior probability and the likelihood function, integrated over μ and Σ.

The resulting joint posterior predictive distribution Pred(Y,$\underline{X}$) may be rewritten as Pred($\underline{Y}$, X_I, X_J), where $\underline{X}$ has been partitioned into subsets X_I and X_J. This predictive distribution is a multivariate student t-distribution (Ando and Kaufman, 1965) with n-1 degrees of freedom, whose mean and variance are:

(21) $E(\tilde{\underline{Y}}, \tilde{X}_I, \tilde{X}_J \mid \underline{Y}, X_I, X_J) = \underline{m}$

where m is the sample mean, i.e. $\underline{m} = \begin{bmatrix} \overline{X}_I \\ \overline{X}_J \\ \overline{Y} \end{bmatrix}$, and

where $\underline{Y}$, X_I, X_J are the Y and X scores observed in the first sample.

(22) and $V(\tilde{\underline{Y}}, \tilde{X}_I, \tilde{X}_J \mid \underline{Y}, X_I, X_J) = \frac{nV}{n-2}$

where V is the sample variance-covariance matrix, and

where n is the number of S's.

Let us now consider the effects of weighting the X's to predict Y. More specifically, if we choose only some of the X's (set X_I) and weight them using a function $f(\underline{X}_I)$,

the loss due to choosing set X_I as some weighting of X, $f(X_I)$, is

$$(23) \quad L = (\tilde{\underline{Y}} - f(\tilde{\underline{X}}_I))'(\tilde{\underline{Y}} - f(\tilde{\underline{X}}_I)) = \sum_{i=1}^{n^*} (\tilde{\underline{Y}} - \tilde{\underline{Y}}_I)^2$$

Lindley attempts to include the cost of the various predictor variables in the above loss function. Denoting the loss due to the cost of gathering information on a set I of the X variables as C_I, the overall loss from predicting Y to be $f(X_I)$ is:

$$(24) \quad L = \sum (\tilde{Y}_i - f(\tilde{\underline{X}}_I))^2 + C_I$$

The problem then is to choose a set I of the predictors and a method of weighting them ($f(\tilde{X}_I)$) such that L is minimized. This is the Bayesian approach.

In determining how to select the best set of variables and determining how to weight them, we need to consider the loss function in conjunction with the set of all possible "decisions" and states-of-the-world. The set of all decisions represents all possible ways of choosing X's and weighting them. The relevant states-of-the-world are the Y scores associated with the chosen X values. This is illustrated in Table 3.

Table 3: Losses Associated With the Consequence of a Decision for a Given State-of-the-World

Decisions:	Predictive Probabilities of State-of-the-World				
	$P(\tilde{Y}\mid\tilde{X}_1)$	$P(\tilde{Y}\mid\tilde{X}_2)$	...	$P(\tilde{Y}\mid\tilde{X}_I)$	...
$f(\tilde{X}_1)$					
$f(\tilde{X}_2)$					
. . . . .					
$f(\tilde{X}_I)$					
. . . . .					

where $f(\tilde{X}_I)$ is the decision to select variables $\tilde{X}_I$ and weight them in some manner, and

where $P(\tilde{Y}\mid\tilde{X}_I)$ is the predictive distribution of Y, given a selected subset of variables $\tilde{X}_I$.

For each decision, one may compute the expected loss for that decision by averaging the loss over the predicted distribution. In order to minimize the expected loss, one

selects the subset which has the smallest expected loss.

In order to find the loss incurred by selecting a particular set of X's (X_I), one must consider the following:

$$(25) \quad L\,[(\tilde{Y}-f(\tilde{X}_I))^2 | X_I] + C_I$$

where $f(\tilde{X}_I)$ is the decision to select variables X_I, and

where C_I is the cost of predictors X_I.

This expression is the loss incurred by deciding to select a particular set of X's and thereby erring in the prediction of Y scores (psychometric loss) plus the non-psychometric loss.

The expected loss for a particular set of X's (X_I) is expressed as:

$$(26) \quad EL\,[(\tilde{Y}-f(\tilde{X}_I)^2 | X_I] + C_I$$

This expected loss is the average squared difference between the actual Y's and the predicted Y's, given the particular subset of predictor variables X_I, plus the non-psychometric cost of the predictors.

In order to minimize this expected loss, one may utilize the fact that the mean squared error is least about the mean. As stated previously, the $\tilde{Y}$, the $\tilde{X}_I$'s and the $\tilde{X}_J$'s jointly have a multivariate t-distribution. From this joint distribution of $\tilde{Y}$, $\tilde{X}_I$, and $\tilde{X}_J$ we need to consider the conditional predictive distribution of $\tilde{Y}$, given $\tilde{X}_I$. The form of this conditional distribution is also a multivariate

t-distribution which can be clearly seen by considering an example.

Let us consider the simple case where $\tilde{X}_I$ and $\tilde{X}_J$ each contain one variable, $\tilde{X}_1$ and $\tilde{X}_2$, respectively. We may partition the mean vector and covariance matrix as follows:

$$\underline{m}_1 = \begin{bmatrix} \bar{\tilde{Y}} \\ \bar{\tilde{X}}_1 \end{bmatrix} \qquad V_{11} = \begin{bmatrix} SSy & SP_{yx_1} \\ & SSX_1 \end{bmatrix}$$

$$m_2 = \bar{\tilde{x}}_2 = [\bar{\tilde{x}}_2] \qquad V_{12} = \begin{bmatrix} SP_{yx_2} \\ SS_{x_1x_2} \end{bmatrix}$$

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \qquad V_{22} = [SS_{x_2}]$$

The marginal distrubtion of $\tilde{Y}$ and $\tilde{X}_1$ can be shown to be (De Groot, 1970) a multivariate t-distribution with n degrees of freedom, with mean vector $\underline{m}_1$ and covariance matrix V_{11}.

The conditional distribution of $\tilde{Y}$, given $\tilde{X}_1$, according to De Groot (1970), is a t-distribution. The conditional distribution has (n+1) degrees of freedom. The mean vector is equal to:

$$(27) \quad \bar{Y} + SP_{yx_1}/SS_{x_1}$$

The variance of this conditional distribution is:

$$(28) \quad \frac{n + \sum(\tilde{x}_1 - \bar{\tilde{x}}_1)^2}{n + SS_{x_1}} \quad (SS_y - \frac{SP^2_{yx_1}}{SS_{x_1}})$$

Since $\sum(\tilde{X}_1-\tilde{\bar{X}}_1)^2 = SS_{x_1}$, the above expression is equal to:

$$(29) \quad (SS_y - \frac{SP^2_{yx_1}}{SS_{x_1}}) = RSS_1.$$

This expression is the R.S.S. resulting from using variable X_1 to predict Y. The values in this expression are obtained from the data in the first sample. The variance of the conditional distribution of $(\tilde{Y}|\tilde{X}_1)$ is also the expected loss resulting from using $E(\tilde{Y}|\tilde{X}_1)$ to predict Y from $\tilde{X}_1$.

Therefore, the minimum expected loss in using variable X_1 is equal to: $R.S.S._1 + C_1$

where $R.S.S._1$ is the residual sum of squares using variable X_1, obtained in the first sample, and

where C_1 is the cost of using variable X_1.

The minimum expected loss in using variable X_2 is equal to: $R.S.S._2 + C_2$

where $R.S.S._2$ is the residual sum of squares using variable X_2, obtained in the first sample, and

where C_2 is the cost of using variable X_2.

This result may be generalized to the cases of more than one variable in set X_I and more than one variable in set X_J, as cited in De Groot (1970) and Lindley (1968).

Lindley's expression for the expected loss is actually

an opportunity loss, i.e. the loss incurred by using the X_I variables minus the minimal loss incurred by using all of the variables.

Lindley denotes this opportunity loss as R(J:I), where:

(30) $R(J:I) = R_I - RSS$

where R_I is the residual sum of squares due to using subset I, and

where RSS is the residual sum of squares using all the variables.

This quantity can vary from 0 (when R_I = R.S.S.) to R_I (when R.S.S. = 0). Thus, as the difference between the total R.S.S. and the subset R.S.S. (R_I) increases, the psychometric cost increases.

Since the loss is the sum of the psychometric loss and the monetary loss, one may represent this loss as:

(31) R(J:I) + non-psychometric cost.

Let us now consider the determination of the monetary cost of subsets of variables. This will be made clearer by investigating an example. Suppose that there are 3 variables in a set (k = 3) and we wish to select a subset of p variables (p 3) from this larger set. Let us say that for a particular subset size (i.e. number of predictors), there is a corresponding residual sum of squares. That is,

for $p = 3$ $\quad R_I = 10$

$p = 2$ $\quad R_I = 12$

$p = 1$ $\quad R_I = 14$

$p = 0$ $\quad R_I = 20$

We may consider different measures of the non-psychometric cost and observe how they are related to the psychometric cost.

Table 4: Psychometric and Non-Psychometric Costs

	Psychometric Cost		Non-Psychometric Cost					
		↓	1	2	3	4	5	6
p	R_I	R_I-RSS	p·RSS	$\frac{p}{k}$ · RSS	$\frac{p}{k} \cdot \sqrt{RSS}$	$\frac{p}{k} \cdot \sqrt[3]{RSS}$	$\frac{p}{2k} \cdot \sqrt{RSS}$	$\frac{p}{3k}$ · RSS
0	20	10	0	0	0	0	0	0
1	14	4	10	3.3	1.1	.7	.55	1.1
2	12	2	20	6.7	2.2	1.4	1.1	2.2
3	10	0	30	10	3.2	2.2	1.6	3.3

Very clearly, measure 1 indicates that the monetary cost has a much larger scale than the psychometric cost. While measure 2 indicates that the non-psychometric cost has a scale somewhat equivalent to the psychometric cost, this may not reflect the practitioner's assessment. The practitioner often feels that an error in prediction is greater than the non-psychometric loss of gathering additional data. Measures 4 and 5 indicate that the psychometric cost is much smaller in scale than the non-psychometric cost. Therefore, we shall choose measures 3 and 6 as the non-psychometric cost

since they attribute a greater scale to the psychometric cost, which reflects the practitioner's typical assessment of the relationship between the two costs. Hence, we shall let first the monetary cost equal

$$\frac{p}{k} \cdot \sqrt{R.S.S.}$$

Then we shall compare these results to those when the following non-psychometric cost is used:

$$\frac{p}{3k} \cdot R.S.S.$$

Therefore, one attempts to minimize this loss. This expression may also be written as:

(32) $\min_I \; R(J:I) + C_I$

where $R(J:I)$ is the reduction in sum of squares due to selecting $X_I = (RSS_I - RSS)$,
where C_I is the cost of predictors I,
$C_I = \sum C_i$, and
where C_i are the cost of predictors in the subset.

Therefore, the Bayesian approach is intimately related to the utility or cost of the selection of a particular subset of predictors, not merely to the existence of differences between the R.S.S. Nonetheless, Lindley's approach can be shown, under certain circumstances, to be a modification of L.S.E. comparison among all subsets, from an applied

viewpoint. In theory, of course, the two approaches are not so readily reconcilable.

For example, in the Gorman and Toman 2 variable case data (see Appendix B), the optimum solution to the prediction problem is the selection of I predictors which satisfy

$$\min_I [R(J:I) + C_I]$$

Using the data,

$$RSS_{12} = 10.07$$
$$RSS_1 = 12.48$$
$$RSS_2 = 11.23$$

where RSS_{12} is the R.S.S. using both variables,

where RSS_1 is the R.S.S. using only variable 1, and

where RSS_2 is the R.S.S. using only variable 2.

When one uses only variable 1,

$$RSS_1 - RSS_{12} = 12.48 - 10.07 = 2.41.$$

When one uses only variable 2,

$$RSS_2 - RSS_{12} = 11.23 - 10.07 = 1.16$$

Using variable 1 + 2 yields:

$$RSS_{12} - RSS_{12} = 10.07 - 10.07 = 0.$$

Using neither variable 1 nor variable 2 yields:

$$RSS_{12} - 0 = 10.07 - 0 = 10.07.$$

When the non-psychometric cost is equal to $\frac{p}{k}\sqrt{R.S.S.}$:

Using variable 1: $R(1:2) + C_I = 1.16 + 1.6 = 2.76$

Using variable 2: $R(2:1) + C_I = 2.41 + 1.6 = 4.01$

Using both variables: $R(0:1,2) + C_I = 0 + 3.2 = 3.2$

Using neither variable: $R(1,2:0) = 10.07$

Minimizing this loss would lead to using variable 1 as the best subset.

Review of the Literature

Numerous techniques have been published which attempt to select a subset of variables from a larger set of variables. Our study deals only with the Forward Selection, Ridge Regression, and Bayesian selection procedure. The Forward Selection procedure was chosen because it is a popular approach and can be used as the basis for comparison with the latter two approaches which are innovative.

Nonetheless, literature has been cited with respect to the following procedures:

1) Comparison of all possible regressions
2) Forward Selection
3) Backward Selection
4) Combination of Backward and Forward (Stepwise)
5) Factor Analytic

6) Mallows' Cp

7) Ridge Regression

8) Bayesian Selection

Although many of the procedures have not been compared with one another, where comparisons have been investigated, results are reported as well.

Comparison of All Possible Regressions

Comparison of all possible regressions has been investigated by Morgan and Tator (1972), Aitkin (1974), Allen (1971), and Spjotvoll (1972), using F-tests to determine which subset yields the smallest residual sum of squares or mean prediction error. Basically, the approach entails computing 2^k (where k is the number of predictors) multiple regressions, i.e. the multiple regression for each and every combination of predictor variables to ascertain which has the smallest R.S.S. or mean prediction error. However, such comparisons may be undertaken only when the number of predictors k is small. Since the number of all multiple regressions is 2^k, for large k, comparison of all possible regressions is a prodigious task.

Forward Selection

Numerous Forward Selection techniques have been devised to select a subset of predictors from a larger array of predictors. Approaches by Wherry (1940), Dwyer (1945),

Summerfield and Dubin (1951), Davies (1958), and Brownlee (1965) choose one predictor at a time in an experimental sample which provides the largest "incremental validity," i.e. increase in the multiple correlation. The first predictor chosen is the one which has the highest correlation with the criterion. The second predictor selected is the one which, together with the first, gives the highest correlation with the criterion. Additional predictors are chosen in this way until the F-test indicates the addition of variables does not add significantly to the multiple R^2. However, these methods do not take into consideration that the intercorrelation among predictors may mean that a configuration of certain variables may yield a higher multiple correlation. Also, once a variable has been included it can no longer be deleted.

Backward Selection

In the Horst and MacEwan (1960) and Mantel (1970) approaches the investigator initially includes the entire pool of predictors. Predictors are sequentially eliminated which in combination with all the others contributes least in predictive value for the criterion. The second measure eliminated is the one which in combination with the remaining p-1 predictors contributes least in predictive value for the criterion. Dropping one variable at a time, the procedure is continued until the F-test indicates that the deletion of variables significantly reduces the multiple R^2.

Stepwise: Combinations of Backward and Forward Selection

Combinations of the two previous approaches are possible, whereby one may successively add or delete variables, according to some specified criteria. This has been termed Stepwise procedure, developed by Effromyson (1960). The Stepwise procedure starts with the simple correlation matrix and enters into the regression the predictor variable most highly correlated with the criterion. Using the partial correlation coefficients, it selects as the next variable to enter the regression that predictor variable whose partial correlation with the criterion is highest. Given the new regression equation, the procedure determines the contribution that the first variable would have made if the second variable had been entered first and the first variable added second. If the partial F is statistically significant at a specified α level, the first variable is retained. The Stepwise procedure then selects the next variable to be included, that is the most highly partially correlated with the criterion. A new regression is now determined by least-squares. Partial F-tests for the first two variables are computed to determine if they should remain in the regression. This procedure continues till the addition of a variable is non-significant on the basis of the partial F-test, at which point the process terminates.

Comparison of Forward and Backward Selection Techniques

Comparison of Backward and Forward techniques has been undertaken by Lord and Novick (1968), who report that the respective techniques do not always lead to the selection of the same subset of predictor variables. For example, in the case of a set of ten predictors, the Forward Selection procedure may select variables 1, 3 and 10, whereas the Backward Selection procedure might select variables 9, 7, 3 and 6.

Factor Analysis and Principal Component Approaches to Variable Reduction

Factor Analytic approaches do not eliminate or reduce the numbers of predictor variables but rather form linear combinations of the original variables. Horst (1941) derives regression equations in terms of the principal components of the predictors and for an underlying factor analysis model. Herzburg (1969) and Jeffers (1967) have also found the use of a relatively small number of principal components as composite predictors to be quite effective when the number of predictors is large relative to the calibration sample size.

But as stated before, Factor Analytic approaches and some Principal Components procedures necessitate collecting data on all the predictor variables. Therefore, it is

preferable to use approaches which do not require collecting data on the entire set of predictors.

Burket (1964), utilizing 29 predictor variables, compares Forward Selection, Backward Selection, use of largest principal components of the predictor intercorrelation matrix, i.e. the best least-square approximation to the predictor intercorrelation matrix, use of the smallest principal components, i.e. the best lower-rank approximation to the inverse of the intercorrelation matrix, and the use of principal components yielding the highest multiple correlation with the criterion. Weight validities were computed by determining the correlations between predicted and observed scores in the new samples. His results suggest that the method of largest components was theoretically and empirically superior to the other methods. Largest principal components yielded more accurate weight validities, especially in small samples, than did principal components having the highest multiple correlation with the criterion.

Comparison of Factor Analytic and Principal Component With Backward and Forward Selection

Rock, Linn, Evans and Patrick (1970) use a Monte Carlo simulation of data procedure to generate sample correlation matrices of differing sample sizes from each of two population correlation matrices (one having high intercorrelations, the other having low intercorrelations). They

compare four techniques: Forward Selection, Backward Selection, and two Factor Analytic techniques. The first Factor Analytic procedure is the Component Predictor Elimination method, which requires the sequential application of Principal Component analysis to the predictor correlation matrix and elimination of one predictor correlation matrix and elimination of one predictor after each component analysis. The predictor that has the largest loading on the characteristic vector associated with the smallest latent root is eliminated after each successive principal component analysis. The second Factor Analytic method is the Collinearity method, which successively eliminates from the predictor matrix the variable which most closely approximates a linear function of the remaining variables. The variable that has the highest multiple correlation with the remaining P-1 predictor variables is eliminated first, then the variable with the highest multiple correlation with the remaining P-2 predictor variables is eliminated. This procedure is applied successively.

The four selection techniques were used for each of 60 calibration sample correlation matrices. Weight validities were computed by using the raw data for the population correlation matrices as if they were cross-validation samples. Average weight validities for each sample size and population matrix were used to evaluate the selection techniques.

Results indicated that the Forward and Backward Selction procedures did as well or better than the Factor Analytic

approaches. The Forward Selection procedure yielded the best results for the small sample size.

Mallows' Cp

Mallows (1964) suggests utilizing as a criterion the standardized total squared error, which is a function of the residual sum of squares, the number of predictors in a subset, and the sample size.

In mathematical terms, the total squared error is:

$$\sum_{j=1}^{N} (B_j - \hat{B}_j)^2 + \sum_{j=1}^{N} \text{var}\ (Y_j)$$

where B_j = expected value from the true equation,

where $\hat{B}_j$ = expected value from the fitting equation,

where $(B_j - \hat{B}_j)$ = bias at point j, and

where $\Sigma(B_j - \hat{B}_j)^2 = SS_B$.

When the total squared error is standardized, one obtains:

$$\Gamma_p = \frac{SS_{Bp}}{\sigma^2} + \frac{1}{\sigma^2} \sum_{i=1}^{N} \text{var}\ (Y)$$

Since $\text{var}(Y) = p\sigma^2$, the above equation reduces to:

$$\Gamma_p = \frac{SS_{Bp}}{\sigma^2} + p$$

The residual sum of squares (R.S.S.) from a p-term equation has the expectation:

$$E(RSS_p) = SS_B + (N-p)\sigma^2$$

Rewriting this equation for SS_B,

$$SS_B = E(RSS_p) - (N-p)\sigma^2$$

Substituting this expression in the equation for Γ_p yields:

$$\Gamma_p = \frac{E(RSS_p)}{\sigma^2} - (N-p) + p$$

or simply,

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (N-2p),$$

where RSS_p is the residual sum of squares for a subset of p predictors,

where $\hat{\sigma}^2$ is an estimate of σ^2, often the RSS for all predictors, and

where p is the particular set's number of predictors.

Mallows further indicates that in regressions with small bias, i.e. where the sum of squared errors for the p predictors is approximately equal to that for all k predictors, C_p will be nearly equal to p and, together with the magnitude of C_p, these are to be used as criteria for the selection of the subset of predictors.

When the total number of variables k is not too large, the residual sum of squares can be determined for each of the 2^k possible regressions and compared with Cp as the criterion. As k becomes large, alternate schemes have been proposed by Gorman and Toman (1966) and Hocking and Leslie (1967). They both assume that, given a subset of p predictors out of a total of k such predictors, only a few regressions are to be considered superior. Moreover, in order to minimize C_p, it is necessary to minimize RSS_p. They have therefore developed procedures to do this which allow a subset to be identified after having determined residual sum of squares for only a relatively small fraction of possible subsets.

Gorman and Toman (1966) analyze data having 2, 6, and 10 predictor variables in a Forward Selection approach with Mallows' criterion. Their approach, a fractional-factorial technique, evaluates in a forward direction select groups of variables which in combination reduce the size of C_p.

Hocking and Leslie (1967), utilizing Mallows' criterion, systematically delete variables one at a time in Backward Selection manner. They reanalyze the Gorman and Toman data using their technique, leading to the selection of one out of two of the identical subsets which Gorman and Toman select. Although their technique does not isolate several potential subsets of regressions, their approach is more methodical than that of Gorman and Toman.

Ridge Regression

Ridge Regression was developed by Hoerl and Kennard (1970) to minimize the mean square error of the coefficient estimates, by adding a constant to the inverse of the $X'X$ matrix. They utilize a Ridge Trace which portrays the variance of the coefficient estimates. Therefore, the applications of Ridge estimation are twofold: either to produce better (less variant from sample to sample) estimates of the coefficients, or to select a subset of variables which are more stable from sample to sample. While several studies (McDonald and Galarneau, 1975; Wermuth, 1972; Marquadt, 1970) have examined the former application, there as yet has been only one study (Hoerl and Kennard, 1970) which examines the latter.

Comparison of Ridge Regression With Forward and Backward Selection, Utilizing Mallows' Cp

Hoerl and Kennard (1970) reanalyze the Gorman and Toman (1967) data and find that two procedures for the selection of a subset utilizing Mallows' criterion yield a subset whose coefficients are unstable (across k). By analyzing a Ridge trace, graphical representation of the coefficients for a range of constants added to the inverse of the $X'X$ matrix, they select a subset of variables whose coefficients do not change drastically for different values of k.

Comparison of Ridge Regression With Least-Squares

Marquadt and Snee (1975) compare Ridge estimation with Least-Squares estimation for three different sets of data. They tabulate the following measures:

S_e = residual standard error from the estimation data

$R^2 = 1 - s^2_e/S^2_y$

VIF = maximum variance inflation factor

S_p = prediction standard deviation at p prediction points

$= (\sum[\hat{Y}_i - E(\hat{Y}_i)]^2/p)^{1/2}$

They find that the estimation residual error (S_e) increases as k increases, due to bias. Accordingly, the R^2 decreases. However, as k increases the variance inflation decreases. The prediction residual error (S_p), which is a function of the bias and the variances, was found to achieve a minimum for $k > 0$. Furthermore, in two studies (cited by Marquadt and Snee) which used cross-validational samples in addition to calibration samples, Ridge estimation yielded a more realistic selection of variables for the real data.

Guilkey and Murphy (1975) empirically compare least-square estimates with Ridge estimates and a modified type of Ridge estimates. Their results indicate that the Ridge estimates and the modified version of the Ridge estimates have a smaller mean square error in the coefficient estimates than do least-squares estimates.

Bayesian Selection of Predictors

Bayesian selection of predictors has been explored by Lindley (1968). Although several studies (Lindley and Smith, 1972; McDonald and Galarneau, 1975; Wermuth, 1973; Zellner, 1971; Zellner and Chetty, 1966; Raiffa and Schlaiffer, 1961; Theil, 1961; Theil, 1963) have dealt with the estimation of regression coefficients in a multiple regression, scarce empirical work has been done within the Bayesian framework to use the methodology to select a subset of predictors. Lindley (1968) compares his approach with a L.S.E. procedure to show how differential costs would lead to differences in subset selection between the two approaches.

Comparison of L.S.E., Cp, Principal Components, Ridge Regression, and Bayesian Selection

In a dissertation by Wermuth (1973) Least-Squares estimates are compared with all variables. Forward Selection, Backward Selection, Mallows' Cp, regression on principal components, Stein-type estimators (which reduces the size of coefficients uniformly), Ridge Regression, and Bayesian Selection (with diffuse prior) for fixed p-predictors. By incorporating various stopping rules for some of the procedures, Wermuth actually investigates 57 variants of the above procedures. She utilizes two overall data sets, each generating 32 simulated sets for six predictors variables, with 20 observations each. Her criteria are:

1) accuracy of prediction = distance between estimate and expected value of dependent variable.

In mathematical terms,

$$SPE = \frac{\Sigma(\hat{Y} - E(Y))^2}{\sigma^2}$$

$$\text{where } \sigma^2 = \Sigma(e^2)$$

2) accuracy of coefficient estimation = distance between estimates and actual values of coefficients.

Mathematically represented as,

$$SEB = \frac{\Sigma(\hat{B}_i - B_i)^2}{\sigma^2}$$

Wermuth ranks the 57 methods on each of the above criteria based on mean, median, and rank where low values on the criteria indicate less error and are hence desirable.

Her results favor Bayesian approaches, with diffuse prior, which add a constant to the inverse of $X'X$, analogous to the Ridge Regression approach but choosing a different constant value from that of the Ridge estimation procedure. Wermuth also found that the Ridge estimates were superior to the L.S.E. for estimation and prediction, particularly when the predictors were multicollinear. However, she does not use the Ridge Regression to select a subset of predictors, only to revise estimates of the coefficients. Further, the major drawback to the study is that she does not cross-validate her results. As argued previously, the

criterion of importance is the performance of prediction and estimation in future samples. Lastly, some of the methods which she compares are not subset selection techniques but are coefficient estimation procedures. The latter do not obviate the need for collecting predictor variable data.

Summarization of the Statement of Problem

Attempts to select an optimal subset of predictors from a set of predictors are numerous; three such methods have been discussed and chosen for investigation. The Forward Selection procedure has been chosen since it is commonly used by practitioners. The Ridge Selection and Bayesian Selection procedures have been chosen since they are more innovative techniques and have not been duly investigated. Hence there exists a need to compare these diverse procedures in order to determine under what circumstances, if any, one procedure is "better" than another.

Each selection procedure attempts to minimize the value of some criterion. The Forward Selection procedure attempts to minimize the average residual sum of squares in the calibration sample. The Ridge Selection procedure attempts to minimize $E(L^2)$, i.e. the average squared difference between the value of the population B's and the sample $\hat{B}$'s computed in the calibration sample. The Bayesian Selection procedure attempts to minimize the average psychometric plus non-psychometric cost with respect to future

samples. However, in addition to these criteria, the goal of a selection procedure should more relevantly be directed to two other criteria, namely 1) maximizing the average weight validity, i.e. the correlation between Y and $\hat{Y}$ with respect to the entire population, and 2) minimizing the mean squared error, i.e. the average squared distance between $\underline{B}$ and $\hat{\underline{B}}$, i.e. when $\hat{\underline{B}}$'s derived in the calibration sample are applied to the entire population.

Therefore, in summary, it is necessary to determine which procedures yield subsets having 1) the highest average weight validity, 2) the highest average mean squared error validity 3) the lowest average R.S.S., 4) the lowest $E(L^2)$, and 5) the lowest average cost.

Expectations

The following are the expectations of the results:

1) That the Forward Selection procedure will, on the average, yield subsets having the smallest residual sum of squares since this procedure seeks to minimize the R.S.S. in selecting subsets.
2) That the Ridge Selection procedure will, on the average, yield subsets having the smallest mean square error since this procedure seeks to minimize the mean square error in the sample.
3) That the Bayesian Selection procedure will, on the average, yield subsets having the smallest cost since this procedure seeks to minimize the cost in each sample.
4) That, with respect to the average weight validity and the average mean square validity indices, the Ridge procedure will yield higher index validities. When coefficients derived in the sample are applied to population parameters, the Ridge coefficients, which have smaller mean square errors, will be better estimates of population coefficients, particularly when the intercorrelation among the predictors is high.

Method

Monte Carlo simulation of data was used to compare the three variables selection techniques · the five criteria, except that instead of using the R.S.S. as criterion, the adjusted R^2 was used since it takes into account the number of predictors selected relative to the sample size. The formula for the adjusted R^2 is:

$$\bar{R}^2 = 1 - \frac{R.S.S./df}{T.S.S./df}$$

where T.S.S. is the total sum of squares.

This formula may be rewritten in terms of the sample size (N) and the number of predictors (P) selected, as follows:

$$\bar{R}^2 = 1 - \frac{R.S.S./(N-P)}{T.S.S./(N-1)}$$

(As P/N approaches 0, $\bar{R}^2$ approaches R^2.)

The simulation of data was used to generate samples from three specified predictor intercorrelation levels. We have considered the case when the total set of predictors consisted of ten predictors. Intercorrelations from 0 to .33 were operationally defined as low, intercorrelations from .34 to .67 were defined as intermediate or medium, and intercorrelations from .68 to 1.0 were defined as high. Given these three predictor intercorrelation levels, an assortment

of predictor-criterion intercorrelation levels were considered for each predictor intercorrelation level[1], as follows:

When the intercorrelation level among predictors was randomly selected within the "low" range, the following five predictor-criterion intercorrelations were examined in conjunction with low intercorrelations among predictors:

1) when all predictor-criterion correlations were low
2) when all were medium
3) when some (5) were low and some (5) were medium
4) when some (9) were low and some (1) were high
5) when some (9) were medium and some (1) were high

When the predictor intercorrelations were in the intermediate level, the following six predictor-criterion intercorrelation levels were considered in conjunction with intermediate predictor intercorrelations:

1) when all predictor-criterion correlations were medium
2) when all were low
3) when all were high
4) when some (7) were medium and some (3) were high
5) when some (7) were medium and some (3) were low
6) when some (5) were medium, some (4) were low, and some (1) were high

Similarly, when the predictor intercorrelation level was high, the following four predictor-criterion intercorrelation levels were examined in conjunction with the high predictor intercorrelations:

[1] the eigenvalues are reported in Appendix F.

1) when all predictor-criterion correlations were high

2) when all were medium

3) when some (5) were high and some (5) were medium

4) when some (9) were medium and some (1) were low

In total, fifteen populations were used to generate subsequent samples. These populations represent a variety of typical intercorrelations found in educational psychology. These population intercorrelations are also meaningful in the sense that, for example, where the interpredictor correlations were low, the predictor-criterion intercorrelations were not all high. Such a situation is impossible since if all the predictors correlate highly with a criterion, they would perforce not all have low intercorrelations. Similarly, where the interpredictor correlations were high, we did not consider uniformly all low predictor criterion intercorrelations since such a situation is likewise impossible.

We have included for consideration the case when the cost of predictors is P/K $\sqrt{\text{R.S.S.}}$ and when it is equal to P/3K R.S.S. (K is the total # predictors in the set.)

Two hundred samples of N = 25, N = 50, and N = 100 were each drawn at random from the fifteen populations. Subsets of predictors were then selected, using the sample data, based on the three selection techniques: 1) Forward Selection, using a sequential F-test; 2) Ridge Regression with Hoerl and Kennard's criterion of discarding variables whose coefficients are close to zero or change in sign and refining the new subset by augmenting the new intercorrelation

matrix with an explicit value of K; and 3) a Bayesian approach, with Lindley's criterion. By drawing the two hundred samples and computing $\hat{B}$ values based on these selection procedures, we computed the Adjusted R-squared, $E(L^2)$, and the expected cost. The $\hat{B}$'s were subsequently applied to the population to determine: a) the average weight validity; and b) the average mean squared error validity.

Let us consider in detail the analysis that was performed in any one of the generated samples. In order to select a subset of predictors from the set, the three variable selection procedures were applied. Further, the statistics needed to obtain the expected values of each of the five criteria were computed. A computer program was written (see Appendix E) to compute all necessary computations. We shall first consider the Forward Selection procedure:

1) For the Forward Selection procedure, the computer program selected the predictor having the highest correlation with the criterion, adding those predictors which added significantly to the regression sum of squares. Thus a regression model would be expounded:

$$\hat{Y} = \hat{B}_o + \hat{B}_1 X_1 + \ldots . \hat{B}_p X_p$$

where p is the number of predictors in the subset.

The adjusted R^2 was computed and averaged over

the many samples; $E(L^2) = E(\Sigma(\hat{B}_i - B_i)^2)$ over many samples was also computed. The cost = $(R.S.S.)_I + C_1$ was computed and averaged over the many samples. Using calibration sample selection and weighting of predictors, the average weight validity and mean squared error in the population were calculated (Eqs. 2A + 2B).

2) For the Ridge estimation procedure, $\hat{\underline{B}}^*$ were found for the sample regression model:

$$\hat{Y} = \hat{B}_o + \hat{B}_1 X_1 + \ldots . \hat{B}_{10} X_{10}$$

such that $\hat{\underline{B}}^* = (X'X + KI)^{-1} X'Y$ for K = 0.0, .10, etc., i.e., intervals of .1 from 0 to 1. We examined each of the $\hat{B}_i^*$ in the Ridge trace as follows:

Any predictors whose coefficients $\hat{\underline{B}}^*$ are less than .05 or change in sign were discarded. Once variables were discarded, the coefficients of the remaining variables were refined. A value of $K = \frac{p\hat{\sigma}^2}{\underline{B}^*/\underline{B}^*}$ was chosen, yielding new coefficients $\hat{\underline{B}}^{**}$ for the selected subset. The average $\bar{R}^2$, $E(L^2)$, and the average costs were computed. Using calibration sample selection and weighting of

predictors, the average weight validity and mean squared error in the population were computed.

3) The Bayesian approach compares all subsets of variables, using L.S.E. of the coefficients except that the variable of monetary cost was also included. We have used two different Bayesian cost functions. The first (Bayesian 1) sought to minimize the total cost when the non-psychometric cost is equal to $P/K \sqrt{R.S.S.}$ The second (Bayesian 2) sought to minimize the total cost when the non-psychometric cost is equal to P/3K R.S.S.

The average $\bar{R}^2$, $E(L^2)$, and the average costs were calculated. Using calibration sample selection and weighting of predictors, the average weight validity and mean squared error in the population were computed.

Results

For the fifteen intercorrelation levels and for sample size of 25, 50, and 100, Tables 5-11 show the five following criteria compared for each selection procedure: 1) the average $\bar{R}^2$; 2) the average M.S.E.; 3) the average Cost 1 and average Cost 2; 4) the average weight validity; and 5) the average mean square error validity. It should again be mentioned that for the second and third criteria,

lower values are desirable, whereas for the first and last two criteria, higher values are desirable.

Result Trends

With respect to the adjusted R^2, our findings indicate that the Forward Selectionprocedure yielded the smallest $\bar{R}^2$ for every sample size and for every intercorrelation level, although as the sample size increased, the proportionate difference between the four procedures decreased. For example, Table 5A indicates that for sample size = 25, the Ridge & Bayesian procedures tended to yield, on the average (i.e., over the different populations), an average $\bar{R}^2$ 15% higher than that of the Forward procedure (the difference ranged from 0%, when the predictors were moderately correlated and the predictor criterion correlations were high, to 60%, when the predictors were moderately correlated and the predictor-criterion correlations were moderate). At N = 100, the Ridge & Bayesian procedures yielded, on the average, an average $\bar{R}^2$ 2.5% higher than that of the Forward procedure (the difference ranged from 0% to 23%).

With regard to the Mean Square Error, our finding was that as the sample size increased and/or the intercorrelation among predictors increased, the Ridge procedure yielded subsets having a lower M.S.E. than the other procedures. However, for the small sample size, the Forward procedure occasionally yielded a smaller M.S.E. Table 6A indicates

that for sample size = 25, the Ridge procedure yielded, on the average, an average M.S.E. 19% lower than that of the Bayesian 1 procedure. For this sample size the Forward procedure occasionally yielded an M.S.E. over 600% higher and occasionally 71% lower than that of the Ridge procedure. The Bayesian 2 procedure tended to yield an average M.S.E. almost 300% larger than that of the Ridge procedure. At the sample size = 100, the Ridge procedure yielded, on the average, an average M.S.E. 27% smaller than that of the Forward procedure (the difference ranging from -50% to over 200% when the intercorrelations were high). The Bayesian 1 procedure yeilded, on the average, an average M.S.E. 10% larger than that of the Ridge procedure whereas the Bayesian 2 procedure yielded an average M.S.E. over 1000% larger than that of the Ridge procedure.

Concerning the average cost, each respective Bayesian procedure yielded subsets having a lower cost than the other procedures for every sample size and for every intercorrelation level. Table 7A indicates that for sample size = 25, the Bayesian 1 procedure tended to yield, on the average, an average cost 1 37% smaller than that of the Forward procedure and 7% smaller than that of the Ridge procedure. At sample size = 100, the Bayesian 1 procedure yielded, on the average, an average cost 1 2% smaller than that of the Ridge procedure and 2% smaller than that of the Forward procedure. Table 8A similarly indicates that for sample size = 25, the Bayesian 2 procedure

Table 5a: Comparison of the 3 Techniques with respect to $\bar{R}^2$ for the 15 populations for N = 25

	Population														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	.281	.961	.720	.584	.843	.556	.219	.971	.737	.653	.699	.723	.466	.828	.561
Ridge	.311	.975	.786	.612	.877	.582	.250	.974	.791	.805	.781	.816	.513	.914	.682
Bayesian 1	.389	.977	.803	.652	.884	.632	.346	.976	.811	.829	.806	.842	.581	.936	.729
Bayesian 2	.391	.977	.800	.648	.883	.627	.350	.976	.809	.825	.804	.839	.578	.934	.726

Legend

Low	1:	low	interpredictor	correlations,	predictor-criterion	correlations: low
	2:	"	"	"	"	: medium
	3:	"	"	"	"	: some low, some medium
	4:	"	"	"	"	: most low, one high
	5:	"	"	"	"	: most medium, one high
Medium	1:	medium	"	"	"	: medium
	2:	"	"	"	"	: low
	3:	"	"	"	"	: high
	4:	"	"	"	"	: most medium, some high
	5:	"	"	"	"	: most medium, some low
	6:	"	"	"	"	: some medium, some low, one high
High	1:	high	"	"	"	: high
	2:	"	"	"	"	: medium
	3:	"	"	"	"	: some high, some medium
	4:	"	"	"	"	: most medium, one low

Table 5b: Comparison of the 3 Techniques with respect to $\overline{R}^2$ for the 15 populations for N = 50

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	.232	.976	.761	.574	.869	.540	.166	.977	.776	.779	.779	.794	.442	.916	.606
Ridge	.259	.976	.774	.595	.876	.560	.192	.978	.799	.799	.799	.819	.500	.919	.667
Bayesian 1	.289	.976	.778	.609	.876	.574	.237	.997	.803	.804	.805	.824	.523	.925	.680
Bayesian 2	.298	.977	.778	.607	.876	.573	.235	.978	.803	.804	.805	.825	.522	.926	.680

*See legend in Table 5a for description of the respective populations.

Table 5c: Comparison of the 3 Techniques with respect to $\bar{R}^2$ for the 15 populations for N = 100

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	.213	.993	.786	.568	.889	.545	.148	.987	.795	.816	.792	.822	.467	.934	.657
Ridge	.231	.992	.789	.580	.890	.556	.168	.986	.802	.816	.795	.828	.500	.932	.670
Bayesian 1	.239	.993	.789	.582	.891	.560	.182	.987	.803	.818	.797	.830	.507	.934	.673
Bayesian 2	.224	.993	.788	.574	.890	.553	.164	.988	.801	.817	.795	.828	.497	.934	.669

*See legend in Table 5a for description of the respective populations.

Table 6a: Comparison of the 3 Techniques with respect to M.S.E. for the 15 populations for N = 25

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	.272	.070	.361	.161	.208	.397	.321	.051	.423	.894	.570	.770	.999	1.033	1.456
Ridge	.639	.013	.175	.348	.101	.586	1.089	.025	.266	.244	.264	.418	1.200	.166	.786
Bayesian 1	.687	.015	.219	.369	.130	.652	1.231	.035	.330	.317	.335	.557	1.607	.273	1.043
Bayesian 2	2.756	.045	.791	1.495	.428	1.615	2.974	.052	.731	.662	.721	.632	1.848	.232	1.209

*See legend in Table 5a for description of the respective populations.

Table 6b: Comparison of the 3 Techniques with respect to M.S.E. for the 15 populations for N = 50

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	.206	.005	.117	.108	.077	.269	.245	.014	.207	.253	.249	.425	.817	.200	.838
Ridge	.257	.006	.071	.137	.041	.245	.455	.012	.109	.097	.108	.192	.547	.072	.356
Bayesian 1	.255	.005	.078	.137	.051	.256	.441	.015	.125	.113	.132	.251	.630	.102	.428
Bayesian 2	2.021	.033	.569	1.093	.312	1.194	2.146	.037	.531	.473	.530	.472	1.345	.168	.883

*See legend in Table 5a for description of the respective populations.

Table 6c: Comparison of the 3 Techniques with respect to M.S.E. for the 15 populations for N = 100

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	.132	.002	.039	.069	.025	.135	.184	.004	.084	.055	.076	.184	.556	.041	.327
Ridge	.105	.003	.031	.058	.019	.101	.187	.006	.046	.039	.045	.083	.228	.029	.148
Bayesian 1	.112	.002	.035	.061	.021	.108	.190	.006	.056	.044	.052	.107	.252	.036	.187
Bayesian 2	1.827	.028	.500	.981	.272	1.055	1.920	.032	.468	.411	.460	.414	1.196	.145	.787

*See legend in Table 5a for description of the respective populations.

Table 7a: Comparison of the 3 Techniques with respect to Cost 1 for the 15 populations for N = 25

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	16.87	.82	6.12	9.40	3.18	10.01	18.58	.59	5.50	7.66	6.48	4.94	12.17	3.74	10.25
Ridge	14.28	.57	4.64	8.14	2.76	8.74	15.33	.57	4.31	4.02	4.35	3.87	9.77	1.81	6.71
Bayesian 1	13.19	.56	4.35	7.52	2.59	8.01	14.10	.54	4.00	3.78	4.04	3.57	8.96	1.69	6.20

*See legend in Table 5a for description of the respective populations.

Table 7b: Comparison of the 3 Techniques with respect to Cost 1 for the 15 populations for N = 50

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	38.17	1.17	11.53	21.12	6.59	22.53	41.56	1.19	11.21	10.41	11.19	10.10	27.30	4.27	19.30
Ridge	36.29	1.18	11.22	20.33	6.56	21.82	39.15	1.21	10.49	9.64	10.49	9.34	24.42	4.07	16.52
Bayesian 1	34.90	1.17	10.93	19.51	6.40	20.91	37.38	1.18	10.13	9.40	10.18	8.96	23.38	3.97	15.92

*See legend in Table 5a for description of the respective populations.

Table 7c: Comparison of the 3 Techniques with respect to Cost 1 for the 15 populations for N = 100

	Population														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	80.56	2.21	23.80	44.33	13.67	47.40	87.03	2.33	22.27	20.18	22.22	19.82	55.51	8.16	36.17
Ridge	79.93	2.23	23.97	44.23	13.79	47.54	86.14	2.33	22.32	20.38	22.31	19.82	53.33	8.24	35.81
Bayesian 1	78.25	2.20	23.62	43.34	13.64	46.37	83.85	2.31	21.91	20.10	21.97	19.36	51.99	8.15	35.06

*See legend in Table 5a for description of the respective populations.

Table 8a: Comparison of the 3 Techniques with respect to Cost 2 for the 15 populations for N = 25

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	16.89	.54	5.87	9.32	2.86	9.94	18.60	.36	5.31	7.44	6.27	4.78	12.13	3.51	10.13
Ridge	14.58	.24	4.05	7.83	2.18	8.45	15.77	.28	3.73	3.38	3.73	3.30	9.56	1.23	6.23
Bayesian 2	13.36	.22	3.82	7.27	2.05	7.79	14.36	.25	3.50	3.19	3.50	3.08	8.77	1.14	5.79

*See legend in Table 5a for description of the respective populations.

Table 8b: Comparison of the 3 Techniques with respect to Cost 2 for the 15 populations for N = 50

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	38.93	.65	11.48	21.43	6.18	22.96	42.13	.75	11.13	10.21	11.08	9.96	27.69	3.67	19.56
Ridge	40.03	.68	11.17	21.46	6.04	23.25	43.63	.75	10.35	9.37	10.33	9.08	26.34	3.35	17.19
Bayesian 2	36.97	.65	10.88	20.22	5.92	21.79	39.63	.72	10.00	9.16	10.03	8.74	24.65	3.29	16.44

*See legend in Table 5a for description of the respective populations.

Table 8c: Comparison of the 3 Techniques with respect to Cost 2 for the 15 populations for N = 100

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	84.34	1.52	25.30	46.23	13.90	49.99	89.61	1.69	23.40	21.37	23.57	20.59	58.07	7.71	38.88
Ridge	91.18	1.58	25.84	48.93	14.06	53.24	99.52	1.74	23.92	21.71	23.97	20.99	60.87	7.78	40.01
Bayesian 2	84.12	1.52	25.24	46.11	13.88	49.83	89.19	1.69	23.23	21.35	23.47	20.30	56.77	7.69	38.37

*See legend in Table 5a for description of the respective populations.

Table 9a: Comparison of the 3 Techniques with respect to I_1 for the 15 populations for N = 25

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	.263	.941	.668	.808	.814	.657	.262	.977	.778	.519	.672	.811	.607	.760	.522
Ridge	.334	.990	.853	.686	.918	.666	.266	.989	.863	.876	.861	.889	.623	.959	.771
Bayesian 1	.326	.988	.827	.685	.897	.642	.254	.984	.837	.848	.833	.865	.576	.939	.731
Bayesian 2	.328	.991	.833	.684	.909	.639	.253	.989	.844	.859	.844	.867	.578	.952	.734

*See legend in Table 5a for description of the respective populations.

Table 9b: Comparison of the 3 Techniques with respect to I_1 for the 15 populations for N = 50.

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	.378	.996	.894	.863	.935	.776	.368	.994	.893	.870	.862	.902	.691	.959	.735
Ridge	.514	.996	.935	.844	.965	.827	.416	.995	.942	.948	.941	.950	.794	.983	.891
Bayesian 1	.495	.996	.930	.842	.958	.817	.410	.994	.933	.940	.929	.939	.773	.978	.875
Bayesian 2	.484	.997	.929	.846	.962	.810	.407	.996	.933	.941	.930	.940	.764	.981	.871

*See legend in Table 5a for description of the respective populations.

Table 9c: Comparison of the 3 Techniques with respect to I_1 for the 15 populations for N = 100

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	.557	.999	.964	.911	.979	.889	.506	.998	.955	.972	.959	.959	.803	.991	.903
Ridge	.718	.998	.971	.927	.984	.922	.627	.997	.974	.979	.975	.978	.906	.993	.953
Bayesian 1	.693	.999	.968	.923	.982	.916	.598	.997	.969	.977	.971	.973	.897	.992	.942
Bayesian 2	.602	.999	.966	.914	.981	.903	.547	.998	.965	.975	.968	.970	.862	.992	.926

*See legend in Table 5a for description of the respective populations.

Table 10a: Comparison of the 3 Techniques with respect to I_2 for the 15 populations for N = 25

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	.756	.172	.427	.751	.394	.668	.811	.374	.507	.293	.406	.511	.692	.211	.482
Ridge	.602	.545	.601	.600	.582	.609	.610	.537	.602	.594	.598	.623	.623	.601	.622
Bayesian 1	.584	.504	.548	.584	.521	.583	.581	.458	.552	.533	.543	.566	.564	.497	.566
Bayesian 2	.293	.280	.278	.291	.276	.289	.292	.277	.278	.275	.279	.284	.281	.277	.280

*See legend in Table 5a for description of the respective populations.

Table 10b: Comparison of the 3 Techniques with respect to I_2 for the 15 populations for N = 50

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	.812	.758	.702	.822	.654	.766	.868	.695	.682	.611	.626	.676	.752	.610	.633
Ridge	.791	.736	.787	.792	.774	.792	.792	.719	.789	.789	.789	.791	.793	.786	.793
Bayesian 1	.791	.743	.774	.791	.741	.785	.797	.677	.767	.764	.758	.758	.776	.741	.770
Bayesian 2	.399	.389	.387	.398	.379	.390	.404	.384	.383	.384	.380	.380	.387	.383	.383

*See legend in Table 5a for description of the populations.

Table 10c: Comparison of the 3 Techniques with respect to I_2 for the 15 populations for N = 100

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	.876	.896	.874	.885	.854	.872	.905	.878	.840	.874	.950	.830	.829	.875	.826
Ridge	.901	.837	.895	.899	.884	.901	.902	.838	.897	.901	.899	.897	.901	.900	.901
Bayesian 1	.896	.877	.885	.895	.871	.895	.900	.842	.881	.892	.886	.877	.894	.887	.883
Bayesian 2	.442	.450	.440	.444	.434	.442	.452	.443	.434	.442	.434	.433	.435	.444	.425

*See legend in Table 5a for description of the respective populations.

yielded, on the average, an average cost 2 69% smaller than that of the Forward procedure and 8% smaller than that of the Ridge procedure. At sample size = 100, the Bayesian 2 procedure yielded, on the average, an average cost 2 4% smaller than that of the Ridge procedure and 1% smaller than that of the Forward procedure.

With regard to the average weight validity, the Ridge procedure tended to yield subsets having higher average weight validity than the other procedures for all sample sizes and for all intercorrelation levels. However, the Ridge procedure was not substantially different from the Bayesian procedures. Table 9A indicates that for sample size = 25, the Ridge procedure yielded an I_1 ranging from -15% to 69%, averaging about 17%, larger than that of the Forward procedure and about 3% larger than that of the Bayesian procedures. At sample size = 50, the Ridge procedure yielded, on the average an I_1 9% larger than that of the Forward procedure and 1% larger than that of the Bayesian procedures. At sample size = 100, the Ridge procedure yielded, on the average, an I_1 6% larger than that of the Forward procedure (ranging from 0% to 29%) and 2% larger than that of the Bayesian procedures (ranging from 0% to 19%).

Our results indicate that when the sample size is large and/or the intercorrelation among predictors is high, the Ridge procedure produced subsets with higher mean squared error validity than the other procedures. However, for the small sample size the Forward procedure occasionally

yielded a higher I_2. The Ridge procedure, also, was not markedly different from the Bayesian 1 procedure. Table 10A indicates that for sample size = 25, the Ridge procedure yielded, on the average, an I_2 57% greater than that of the Forward procedure only when the intercorrelation among predictors was high. However, when the intercorrelation among predictors was low or medium, the Ridge procedure occasionally yielded an I_2 as high as 300% larger than that of the Forward procedure and occasionally 25% smaller than that of the Forward procedure. The Ridge procedure tended to yield, on the average, an I_2 10% larger than that of the Bayesian 1 procedure whereas the Ridge yielded, on the average, an I_2 over 200% larger than that of the Bayesian 2 procedure. For sample size = 50, the Ridge procedure yielded, on the average, an I_2 19% larger than that of the Forward procedure for the high intercorrelations among predictors, 102% larger than that of the Bayesian 2 procedure, and 3% larger than that of the Bayesian 1 procedure. At sample size = 100, the Ridge procedure yielded, on the average, an I_2 2% larger than that of the Forward procedure (ranging from -7% to 9%), 1% larger than that of the Bayesian 1 procedure, and 89% larger than that of the Bayesian 2 procedure.

Discussion

We shall first summarize the findings in terms of our expectations and then we will discuss these findings as well as offer suggestions for future research.

Summary of Results

Our first expectation was that the Forward Selection procedure would yield the minimum residual sum of squares or Adjusted R^2 since the procedure attempts to minimize the R.S.S. in a sample. Our finding was that the other 3 procedures yielded a higher $\bar{R}^2$ for every sample size and for every intercorrelation level, although as the sample size increased and/or the intercorrelation among predictors increased, the proportionate difference between the four procedures decreased. For example, at sample size = 25 the Ridge & Bayesian procedures yielded, on the average, an average $\bar{R}^2$ 15% higher than that of the Forward procedure but at sample size = 100, the difference decreased to 2.5%. Therefore our first expectation was not confirmed.

Our second expectation was that the Ridge Selection procedure would yield a smaller mean squared error than the other procedures, particularly when the intercorrelation among predictors is high. This was to be achieved by the Ridge procedure since the Ridge approach adds an increment to the main diagonal thereby orthogonalizing the matrix, to produce subsets of predictors whose coefficient estimates

would be closer to population coefficients. Our finding was that as the sample size increased and/or the intercorrelation level increased, the Ridge procedure did indeed yield subsets having a lower M.S.E. than the other three procedures. However, this finding was most evident, as mentioned previously, for the larger sample sizes. For example, at sample size = 100, the Ridge procedure yielded, on the average, an average M.S.E. 27% smaller than that of the Forward procedure. Therefore, this expectation was, to an extent, confirmed.

Our third expectation was that the Bayesian procedure would yield subsets having a lower cost than the other procedures. Our finding was that, in regard to the two different cost functions, each respective Bayesian procedure yielded subsets having a lower cost than the other procedures for every sample size and for every intercorrelation level. For example, the average cost 1 ($R.S.S._I + \frac{P}{k}\sqrt{R.S.S.}$) was, on the average, 37% less when the Bayesian 1 procedure was used rather than when the Forward procedure was used at sample size = 25. The average cost 2 ($R.S.S._I + \frac{P}{3k}$ R.S.S.) was, on the average, 69% less when the Bayesian 2 procedure was used in comparison with the Forward procedure at sample size = 25.

Our fourth expectation was that, with respect to the average weight validity index (I_1) and the average mean square error validity index (I_2), the Ridge procedure would yield subsets having higher validities than the other procedures.

This expectation was based on the reasoning of our second expectation that if the Ridge procedure minimized the mean squared error, its coefficient estimates, when applied to population parameters, would subsequently yield higher validity indices. Our finding was that, except for isolated conditions, the Ridge procedures yielded subsets having a higher weight validity than the other procedures for all sample sizes and for all intercorrelation levels. However, the Ridge procedure was not substantially different from the Bayesian procedures. For example, at sample size = 50 the Ridge procedure yielded an I_1 9% larger than that of the Forward procedure and 1% larger than that of the Bayesian procedures. Our results also indicate that, when the sample size was large and/or the intercorrelation among predictors high, the Ridge procedure produced subsets with higher mean squared error validity than the other procedures. The Ridge procedure was not markedly different from the Bayesian 1 procedure. Further, when the sample size was small, the Forward procedure occasionally yielded a higher I_2. For example, at sample size = 100, the Ridge procedure yielded, on the average, an I_2 2% larger than that of the Forward procedure, 1% larger than that of the Bayesian 1 procedure, and 89% larger than that of the Bayesian 2 procedure. Therefore, our fourth expectation was largely confirmed.

Discussion of the Results

With regard to the residual sum of squares or $\bar{R}^2$, it was surprising that the Bayesian and Ridge Selection procedures yielded subsets having a larger average $\bar{R}^2$ than the Forward procedure which attempts to maximize exactly this criterion. However, this can be explained in a number of ways. First, that the Forward procedure does minimize the R.S.S. for any given number of predictors. Were the other procedures to select the same number of variables for a subset, the R.S.S. of the Forward procedure would be smallest in magnitude. By examining Tables 11a-c, which enumerate the average number of variables in the subsets for the different intercorrelation levels and for the different sample sizes, one may readily observe that the Forward Selection procedure uniformly selected the fewest number of variables. The Ridge procedure, by the same token, consistently selected the greatest number of variables; the Bayesian procedures selected an intermediate number. Thus, although the Ridge and Bayesian procedure reduce the R.S.S., they do so only by the addition of variables to the subset. Despite the fact that this reduction in $\bar{R}^2$ seems appreciable for the smaller sample sizes at low intercorrelation levels, this difference was not significant at the .05 level since the Forward procedure stops adding variables when the F-test no longer yields a significant increase. However, the crucial criteria by which the three

Table 11a: Comparison of the 3 Techniques in terms of the Av. no. of predictors for the 15 populations for N = 25

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	1.2	8.2	3.4	1.8	4.6	1.8	.8	6.2	2.6	3.0	3.0	2.3	1.3	3.8	1.9
Ridge	7.8	9.4	8.4	7.4	8.2	7.9	8.0	7.8	8.2	8.9	8.7	7.8	8.2	9.3	8.2
Bayesian 1	5.6	9.3	7.3	5.5	7.0	5.6	5.7	6.9	6.7	7.8	7.3	6.3	6.3	8.4	6.7
Bayesian 2	5.4	9.8	7.8	5.9	8.1	5.9	5.5	9.1	7.5	8.5	8.1	7.4	6.5	9.3	7.2

*See legend in Table 5a for description of the respective populations.

Table 11b: Comparison of the 3 Techniques in terms of the Av. no. of Predictors for the 15 populations for N = 50

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	1.6	9.7	5.8	2.2	6.8	2.5	1.1	8.0	4.3	6.8	5.7	4.0	1.9	8.1	3.9
Ridge	7.6	9.4	8.5	7.3	8.7	7.9	8.3	8.2	8.6	9.2	9.1	8.5	8.5	9.7	8.7
Bayesian 1	5.1	9.5	7.6	5.0	7.9	5.5	5.0	7.7	7.2	8.4	7.9	6.7	6.1	9.1	7.1
Bayesian 2	3.8	9.9	7.6	4.5	8.3	4.7	3.6	9.0	7.2	8.5	8.0	6.8	5.4	9.5	6.8

*See legend in Table 5a for description of the respective populations.

Table 11c: Comparison of the 3 Techniques in terms of the Average no. of Predictors for the 15 populations for N = 100

	Population*														
	Low					Medium						High			
	1	2	3	4	5	1	2	3	4	5	6	1	2	3	4
Forward	2.5	10.0	6.9	2.9	8.3	3.5	1.5	9.0	6.2	8.5	7.6	5.8	3.0	9.5	5.9
Ridge	7.1	9.5	8.6	6.9	9.0	7.6	7.8	8.4	8.7	9.3	9.2	8.5	8.6	9.9	9.0
Bayesian 1	5.0	9.8	7.7	5.1	8.7	5.5	4.7	8.4	7.5	8.9	8.5	7.1	6.4	9.7	7.6
Bayesian 2	2.9	10.0	7.3	3.4	8.6	4.1	2.2	9.2	7.1	8.7	8.2	6.6	4.6	9.7	6.7

*See legend in Table 5a for description of the respective populations.

procedures are to be evaluated are the cross-validation indices.

Concerning the mean squared error it was not contrary to expectation that the Ridge procedure should yield subsets having a lower average M.S.E., especially for higher intercorrelations among predictors. However, our findings indicated that for small sample sizes and/or low intercorrelations among predictors, the Ridge procedure was not as effective in minimizing the M.S.E. Therefore, it remains to be determined how one can minimize this criterion under these conditions. In all fairness it should be mentioned that for the smaller sample sizes, the ratio of sample size to the number of predictors was extremely low. The practitioner is usually forewarned that such small ratios are not statistically judicious. Had the ratio been consistently higher perhaps then the Ridge procedure would have uniformly yielded lower M.S.E. Nonetheless where larger ratios are unavailable in real-life situations, it is possible that other variations of the Ridge technique used here might yield a lower M.S.E. This is clearly an area for future inquiry.

By comparison with the two other techniques which purportedly optimize a specific criterion, the Bayesian procedures are the only procedures which, without exception, succeed empirically in optimizing the criterion which that approach seeks to achieve. It is interesting to note that for the smaller sample sizes the Ridge procedure tended

to yield subsets having a lower average cost (particularly with the first cost function) than the Forward Selection procedure. Thus, despite the fact that the Ridge procedure selected more variables in its subsets, the concomitant reduction of the R.S.S. probably reduced the overall cost.

The finding which confirmed the expectation that the Ridge procedure yielded subsets having a higher average weight validity index than the other procedures clearly indicates the appropriateness of this approach to the problem of variable selection both for researchers and for practitioners. Similarly, the finding that the Bayesian procedures yielded subsets having a higher average weight validity index than the traditional Forward Selection procedure also suggests the meritorious qualities of this approach as an alternative to the traditional procedure. At present, statistical programming packages, which are expedient for the researcher and practitioner, only utilize the traditional Forward, Backward, or Stepwise Selection procedures; however, with time, perhaps other selection procedures will become available.

The finding that the Ridge procedure yielded subsets having higher average mean squared error validity than the other procedures when the sample size was large and/or when the intercorrelation was high, probably is due to the inverse relationship between the expected M.S.E. and I_2. As the difference between the coefficient estimates and population coefficients decreases in the sample, the expected M.S.E.

decreases and I_2 increases. When the M.S.E. was low, the average mean squared error validity tended to be high. The Bayesian1procedure yielded higher I_2 than the Forward procedure when the sample size was large and/or the intercorrelation high which also corroborates the inverse relationship between M.S.E. and I_2. The findings with Bayesian 2 bear this out, too.

However, despite the small sample size to number of predictors ratio which, as mentioned earlier, statistically is injudicious, it would be helpful to develop other possible variations of the Ridge technique to yield higher I_2 for small sample sizes and low intercorrelations. This also was suggested previously.

Implications of the Findings and Suggestions for Future Research

The findings clearly indicate the useful properties of two innovative approaches to variable selection: namely the Ridge Regression Selection procedure and the Bayesian Selection procedure. Both procedures yield a smaller R.S.S. than the Forward approach and a lower M.S.E. when the sample size is large relative to the initial number of predictor variables. The Bayesian procedure uniformly yields a lower cost than the Forward procedure. Lastly, both procedures yield a higher average weight validity and a higher mean square error validity (when the sample size is large).

Although some suggestions for future research have

been included before, we shall reiterate a couple of suggestions and recommend others.

First, it would be worthwhile to explore other variations of the Ridge technique to determine whether they yield subsets having a lower M.S.E. for all intercorrelation levels and even for small sample sizes. For example one may compare the Ridge Selection approach which discards variables, without refining the new subset, with the approach used in this study. Or, one may examine the effects of a more stringent discarding of variables on the part of the Ridge procedure. For example, what would the effects be were one to delete variables whose ridge trace indicated a slope greater than some fixed value? Or yet, what would the effects be in adding different values of K (instead of a constant) in the diagonal to the different predictors?

Second, one may investigate other Bayesian Selection procedures to determine whether they would yield higher weight validity and mean squared error validity than those obtained in this document. For example, one may consider a Bayesian procedure which uses a different cost function from those used in this study. Practical cases might necessitate the formulation of a cost function which relates the psychometric cost with the non-psychometric cost in some different manner to that used in this study.

Third, it would be enlightening to compare the three techniques for a real-life data situation in which the inter-

correlation structure is different from those considered here.

Fourth, comparison of variable selection techniques could be undertaken for other criteria, other than those used here, which are most appropriate to a particular real-world situation or research interest.

Fifth, implementation of the Forward Selection procedure could consider the addition of variables for alpha levels greater than .05.

APPENDIX A

Gorman and Toman 10 Variable Data

N=36

Matrix of Simple Correlation Coefficients, r

	a	b	c	d	e	f	g	h	j	k	y
a	1.00										
b	-0.04	1.00									
c	0.51	-0.00	1.00								
d	0.12	-0.16	0.00	1.00							
e	-0.71	0.06	-0.59	-0.07	1.00						
f	-0.87	0.09	-0.65	-0.09	0.84	1.00					
g	-0.09	0.24	-0.02	0.03	0.38	0.13	1.00				
h	-0.00	0.01	0.34	0.08	-0.36	-0.20	-0.48	1.00			
j	-0.04	0.09	-0.08	0.02	-0.14	0.04	0.07	-0.18	1.00		
k	-0.36	-0.30	-0.44	-0.09	0.54	0.45	0.40	-0.46	0.05	1.00	
y	-0.81	-0.10	-0.63	-0.10	0.56	0.81	0.04	0.06	0.16	0.45	1.00

APPENDIX B

Two-Variable Case (Gorman and Toman data)

X_1	X_2	Y	
-1.0	- .5	1.0	
-1.0	- .5	- .4	$y = .94 + .84\ X_1 + 1.55\ X_2$
-1.0	- .5	- .1	$y = .94 + 2.01\ X_1$
-1.0	- .5	-1.9	$y = .94 + 2.57\ X_2$
-1.0	-1.0	-1.4	
-1.0	-1.0	-1.8	$r_{12}^2 = .90$ $\quad \sigma_1^2 = .24$
-1.0	-1.0	-2.4	$\sigma_2^2 = .28$
-1.0	-1.0	- .6	$RSS_{12} = 10.07$ $\quad MSR_{12} = .775$
1.0	1.0	3.5	$RSS_1 = 12.48$ $\quad MSR_1 = .89$
1.0	1.0	2.9	$RSS_2 = 11.23$ $\quad MSR_2 = .80$
1.0	1.0	2.8	$R_{y.1}^2 = .87$
1.0	1.0	3.8	$R_{y.1}^2 = .84$
1.0	.5	3.8	$R_{y.2}^2 = .85$ $\quad$ MSR = mean square residual (RSS/df)
1.0	.5	2.9	
1.0	.5	2.1	
1.0	.5	1.8	

APPENDIX C

Derivation of Ridge Estimators

Letting $\dot{B}$ be any estimate of B, one may denote the residual sum of squares in the sample as:

$$\emptyset = (\underline{Y}-\underline{X}\dot{\underline{B}})'(\underline{Y}-\underline{X}\dot{\underline{B}})$$

This may be rewritten as:

$$= (\underline{Y}-\underline{X}\hat{\underline{B}})'(\underline{Y}-\underline{X}\hat{\underline{B}}) + (\dot{\underline{B}}-\hat{\underline{B}})'X'X(\dot{\underline{B}}-\hat{\underline{B}}),$$

where the first term on the right is the L.S.E. residual sum of squares (R.S.S.) resulting from the use of L.S. coefficients, and the second term is the additional R.S.S. incurred when using biased estimates of B.

In the attempt to reduce the sampling variance of our estimates, we would like to decrease the absolute value of the computed estimates. Therefore, let us find $\dot{B}$ such that its squared length $(\dot{\underline{B}}'\dot{\underline{B}})$ is equal to a minimum, subject to the constraint that the second term of the residual sum of square is fixed at some constant.

Therefore, to minimize the length of the regression vector $\dot{\underline{B}}'\dot{\underline{B}}$, subject to the constraint,

$$(\dot{\underline{B}}-\hat{\underline{B}})'X'X(\dot{\underline{B}}-\hat{\underline{B}}) = \emptyset_o,$$

one minimizes the function

$$F = \underline{\dot{B}}'\underline{\dot{B}} + (1/k) \quad (\underline{\dot{B}}-\underline{\hat{B}})'X'X(\underline{\dot{B}}-\underline{\hat{B}})-\phi_o$$

where 1/k is the Lagrangian multiplier.

By taking the derivative, with respect to $\dot{B}$, and setting it equal to 0, one obtains

$$\frac{\partial F}{\partial B} = 2\underline{\dot{B}} + (1/k) \; [2(X'X)\underline{\dot{B}}-2(X'X)\underline{\hat{B}}] = 0.$$

Rearranging terms, one derives

$$\underline{\dot{B}} = \underline{\hat{B}}^* = [X'X + kI]^{-1}X'Y$$

The second term is called the Ridge estimator.

APPENDIX D

The R.E. ($\hat{\underline{B}}^*$) can be related to L.S.E. ($\hat{\underline{B}}$) by substituting $X'X + kI$ for $X'X$, as follows:

$$\hat{\underline{B}}^* = [X'X + kI]^{-1}X'Y$$
$$= w\, X'Y$$

where $w = [X'X + kI]^{-1}$

In an alternate form,

$$\hat{\underline{B}}^* = [I + k(X'X)^{-1}]^{-1}\hat{\underline{B}}$$
$$= z\, \hat{\underline{B}}$$

where $z = [I + k(X'X)^{-1}]^{-1}$.

The eigenvalues of w (ζ_i (w) for i = 1, 2, ... p) and z (ζ_i (z) for i = 1, 2, ... p) are then:

$$\zeta_i(w) = 1/(\lambda_i + k)$$
$$\zeta_i(z) = \lambda_i/(\lambda_i + k)$$

where λ_i is an eigenvalue of $X'X$.

The above may readily be shown, since:

$$w = (X'X + kI)^{-1}$$
$$\zeta_i(w) = \zeta_i(X'X + kI)^{-1}$$
$$= (\lambda_i + k)^{-1} = 1/(\lambda_i + k)$$

and similarly,

$$z = [I + k(X'X)^{-1}]^{-1}$$

$$\lambda_i(z) = \lambda_i[I + k(X'X)^{-1}]^{-1}$$

$$= [1 + k/\lambda_i] = \left[\frac{\lambda_i + k}{i}\right]^{-1} = \frac{\lambda_i}{\lambda_i + k}$$

Thus, the eigenvalues of w and z can be shown to be a function of λ_i and k, i.e. R.E. are related to L.S.E.

Using this relationship between the eigenvalues of R.E. and those of L.S.E., it can be shown (Hoerl and Kennard, 1970) that the "size" of the R.E. is never greater than the size of the L.S.E. More specifically, it can be shown that $(\hat{\underline{B}}^*)'(\hat{\underline{B}}^*) \leq \hat{\underline{B}}'\hat{\underline{B}}$, i.e. the squared Ridge regression estimates are smaller than the squared L.S.E. for $k \neq 0$. This can be proven in the following manner:

Let A be the matrix of order n, λ_i its eigenvalues, and N(A) its norm. Then,

$$(26) \quad N(A) = \sqrt{\sum_{j=1}^{n} a_{ij}^2}$$

Rewriting N(A) in terms of its eigenvalues yields:

$$(27) \quad N(A) = [\sum_{i=1}^{n} \lambda_i]^{1/2} \text{ and } N(A^{-1}) = \sum[1/\lambda_i^2]^{1/2}$$

For the matrix C = A + kI, the norm of C is:

$$(28) \quad N(C) = [\sum_{i=1}^{n} (\lambda_i+k)^2]^{1/2} \text{ and } N(C^{-1}) = \left[\sum \frac{1}{(\lambda_i+k)^2}\right]^{1/2}$$

By Minkowski's inequality (cited in Hardy, Littlewood, and Polya (1934)) for $N > 1$,

$$(29)\quad [\Sigma(\lambda_i + k)^2]^{1/2} \leq [\Sigma \lambda_i^2]^{1/2} + kn^{1/2}$$

and also

$$[\Sigma(\lambda_i + k)^{-2}]^{-1/2} \geq [\Sigma \lambda_i^{-2}]^{-1/2} + kn^{-1/2}$$

By taking the inverse of each side of the equation, the direction of the inequality changes such that,

$$\left[\Sigma \frac{1}{(\lambda_i + k)^2}\right]^{1/2} \leq \frac{1}{[\Sigma^{1/\lambda^2}]^{-1/2} + kn^{-1/2}}$$

$$s = [\Sigma 1/(\lambda_i + k)^2]^{1/2} \leq \frac{1}{[\Sigma 1/\lambda^2]^{-1/2}} = T$$

Since $s \leq T$,

$$s\ X'Y \leq T\ X'Y,$$

squaring both sides of the inequality yields,

$$\hat{\underline{B}}^*{}'\hat{\underline{B}}^* = (s\ X'Y)'s'X'Y \leq (T\ X'Y)'T\ X'Y = \hat{\underline{B}}'\hat{\underline{B}},$$

which is what we set out to prove.

Therefore, since $(\hat{\underline{B}}^*)'(\hat{\underline{B}}^*) \leq \underline{B}'\underline{B}$, taking the expectation of these values does not change the relationship; that is, $E(\underline{B}^*)'(\underline{B}^*) \leq E(\hat{\underline{B}}'\hat{\underline{B}})$.

Recalling that,

$$E\ (L^2) = E(\hat{\underline{B}}'\hat{\underline{B}}) - \underline{B}'\underline{B}, \quad \text{and}$$
$$E(L^{*2}) = E(\hat{\underline{B}}^*{}'\hat{\underline{B}}^*) - \underline{B}'\underline{B}$$
$$E(L^{*2}) = E(\hat{\underline{B}}^*{}'\hat{\underline{B}}^*) - \underline{B}'\underline{B} \leq E(\hat{\underline{B}}'\hat{\underline{B}}) - \underline{B}'\underline{B} = E(L^2),$$

which is what we set out to prove.

APPENDIX E :

Computer Program

```
FORTRAN IV G LEVEL 21          MAIN
0001      DIMENSION SIGM(66),SGMXX(66),SGMXX1(66),TA(11,25),A(25,11)
0002      DIMENSION FB(11),JX(10),JH(11),DM(25,11),MX(11),EB(11),REB(11)
0003      DIMENSION TAA(11,11),XY(10),B(11),RVEC(25,11),ICKV(11),VAS(66)
0004      DIMENSION RFB(11,1),IIH(11),TDM(11,25),TXX(11,11)
0005      DIMENSION ISGXX(11,11),VA(66),AA(11,11),YA(66),RINI(9)
0006      DIMENSION C(11,11),INU(11),EE(11),KAA(11,11)
0007      DIMENSION BETA(11),ETA(11,11),B1(11),RB(11),WKVEC(11),XA(66)
0008      DOUBLE PRECISION SEED,ASEED
0009      REAL ISGXX,MX,NOR
0010      DO 502 IG=1,17
C READ SIGM
0011      READ(5,1000) (SIGM(I),I=1,66)
0012 1000 FORMAT(11F5.0)
0013      NUM=200
0014      SEED=.333333333333D0
0015      NK=25
0016      NI=10
0017      NV=NI+1
0018      N1=(NI+1)*NI/2
0019      N2=N1+NV
C SETUP ALL THE COUNTERS.
0020      AKSSE=0.
0021      AKSSR=0.
0022      AMSE=0.
0023      ACVC=0.
0024      AMSC=0.
0025      ACOST=0.
0026      ARMSE=0.
0027      ARCOST=0.
0028      AKCVC=0.
0029      ARMSC=0.
0030      ABRSS=0.
0031      ABMSE=0.
0032      ABCOST=0.
0033      ABCVC=0.
0034      ABMSC=0.
0035      ABRSY=0.
0036      ABMSY=0.
0037      ABCOSY=0.
0038      ABCVY=0.
0039      ABMSY=0.
0040      NOR=0.
```

```
0041                FACOS=0.
0042                RACOS=0.
0043                BACOS=0.
0044                GIF=0.
0045                GIR=0.
0046                GIY=0.
0047                GIB=0.
0048                FIF=0.
0049                RIR=0.
0050                CIB=0.
0051                AIB=0.
                C STARTING POINT FOR GGMRN.
0052                ASEED=SEED
0053                WRITE(6,950) NR
0054          950   FORMAT(10X,'SAMPLE SIZE IS',1X,I3/)
0055                WRITE(6,1001) (SIGM(I),I=1,N2)
0056          1001  FORMAT(10X,'SIGM MATRIX'//6(10X,11(F5.3,1X)/))
              C CALCULATE POP. BETA AND MULTIPLE R.
              C INVERT SIGM MATRIX.
0057                DO 10 I=1,N1
0058          10    SGMXX(I)=SIGM(I)
0059                KI=6
0060                CALL LINV1P(SGMXX,NI,VA,KI,D1,D2,IER)
0061                K=0
0062                DO 11 I=1,NI
0063                DO 11 J=1,I
0064                K=K+1
0065                ISGXX(I,J)=VA(K)
0066             11 ISGXX(J,I)=VA(K)
0067                R=0.0
0068                DO 20 I=1,NI
0069                BETA(I)=0.0
0070                DO 20 J=1,NI
0071                JD=J+N1
0072          20    BETA(I)=BETA(I)+SIGM(JD)*ISGXX(I,J)
0073                DO 21 I=1,NI
0074                II=N1+I
0075             21 R=R+SIGM(II)*BETA(I)
              C GENERATE RANDOM VECTORS.
              C SAVE SIGM IN SGMXX.
```

The random number generator GGNRM is from IMSL.

```
0076            DO 15 I=1,N2
0077        15  SGMXX1(I)=SIGM(1)
            C ***********************************
0078            CALL GGNRM(SEED,NR,NV,SGMXX1,NR,RVEC,WKVEC,IER2)
            C ***********************************
0079            GO TO 1
0080        2    CALL GGNRM1(SEED,NR,NV,SGMXX1,NR,RVEC,WKVEC,IER2)
            CSETUP MATRIX A,SYMMETRIC STORAGE MODE,CORRECTED SS AND CROSS PRODUCTS
            C MEAN CORRECTION.
0081        1   DO 16 I=1,NV
0082            MX(I)=0.0
0083            ZNR=NR
0084            DO 17 J=1,NR
0085        17    MX(I)=MX(I)+RVEC(J,I)
0086        16  MX(I)=MX(I)/ZNR
            CSUBTRACT MEANS AND FORM TRANSPOSE
0087            DO 18 J=1,NV
0088            DO 18   I=1,NR
0089            A(I,J)=RVEC(I,J)-MX(J)
0090        18  TA(J,I)=A(I,J)
0091            CALL VMULFF(TA,A,NV,NR,NV,NV,NR,TAA,NV,IER3)
0092            DO 19 I=1,NV
0093            DO 19 J=1,NV
0094            RAA(I,J)=TAA(I,J)/SQRT(TAA(I,I)*TAA(J,J))
0095        19  C(1,I,J)=TAA(I,J)
            CONVERT TAA INTO VECTOR FORM AND STORE IN VA AND VA1
0096            K=1
0097            DO 23 I=1,NV
0098            DO 23 J=1,I
0099            VA(K)=TAA(I,J)
0100            VAS(K)=TAA(I,J)
0101        23  K=K+1
0102            MN=NOR+1
            C ***********************************
            C L.S.EST. OF BETA AND RSS.
            C ***********************************
0103            DO 25 I=1,NI
0104        25  JX(1)=1
0105            IH(1)=0
0106            ICPT=0
0107            CALL RLFORC(VA,NI,NR,.5,.5,JX,IH,6,IOPT,IER)
```

```
0108                DO 30 I=1,NI
0109          30    BBETA(I,1)=B(I)
                  STORE RSS IN RSSAL
0110                RSSAL=VA(N2)
                  C ************************************
                  C FORWARD SELECTION
                  C ************************************
0111                IIH(1)=0
0112                IOPT=0
0113                CALL RLSTEP(VAS,NI,NR,.05,.50,IIH,FB,IOPT,IER)
                  STORE RKSS IN RSSFW
0114                RSSFW=VAS(N2)
                  COUNT NUMBER OF VARIABLES
0115                IP=0
0116                DO 35 I=1,NI
0117          35    IP=IP+IIH(I)+1
0118                KP=IP/2
0119                CALL MESS(BETA,FB,RSSAL,RSSFW,KP,NI,N1,N2,SIGM,R,ZMSE,COST,CVC,
                   1ZMSC,FCOS,NR,RAA)
0120                QP=KP
0121                QIF=QIF+QP
0122                FIF=FIF+QP*QP
0123                ARSSF=ARSSF+RSSFW
0124                AMSE=AMSE+ZMSE
0125                ACOST=ACOST+COST
0126                ACVC=ACVC+CVC
0127                AMSC=AMSC+ZMSC
0128                FACOS=FACOS+FCOS
                  C ************************************************************
                  C * RIDGE REGRESSION.                                        *
                  C ************************************************************
                  C MAKEUP X'Y.
0129                DO 120 I=1,NI
0130                XY(I)=0.0
0131                DO 120 J=1,NR
0132          120   XY(I)=XY(I)+RVEC(J,I)*RVEC(J,NV)
0133                DO 130 J=1,NV
0134                DO 130 I=1,NR
0135                A(I,J)=RVEC(I,J)
0136          130   TA(J,I)=RVEC(I,J)
0137                CALL VMULFF(TA,A,NV,NR,NI,NV,NR,TXX,NV,IER)
0138                RK=0.
0139                DO 100 KKK=1,10
0140                RK=RK+.1
0141                K=1
```

```
0142                    DO 135 I=1,NI
0143                    DO 135 J=1,I
0144                XA(K)=TXX(I,J)
0145          135   K=K+1
0146                II=0
0147                DO 102 I=1,NI
0148                II=II+I
0149          102   XA(II)=XA(II)+RK
0150                DO 3000 I=1,NI
0151          3000  RRB(I,1)=XY(I)
0152                CALL LEQT1P(XA,1,NI,RRB,NV,6,D1,D2,IER)
0153                K=KKK+1
0154                DO 103 J=1,NI
0155          103   EBETA(J,K)=RRB(J,1)
0156          100   CONTINUE
              C CHECK EBETA.
0157                IK=1
0158                DO 200 I=1,NI
0159                DO 201 K=1,11
0160                DT=ABS(EBETA(I,K))
0161                IF(DT.LT..05) GO TO 202
0162                IF (EBETA(I,1)*EBETA(I,K).LT.0.) GO TO 202
0163          201   CONTINUE
0164                ICRV(I)=1
0165                GO TO 200
0166          202   ICRV(I)=0
0167          200   CONTINUE
0168                ICRV(NV)=1
              C SETUP MATRIX AA AND COLLAPSE TAA AND RVEC.
0169                IK=0
0170                DO 300 I=1,NV
0171                ITV=ICRV(I)
0172                IF(ITV.EQ.0) GO TO 300
0173                IK=IK+1
0174                RB(IK)=XY(I)
0175                B1(IK)=XY(I)
0176                DO 301 II=1,NR
0177                DM(II,IK)=RVEC(II,I)
0178          301   TDM(IK,II)=DM(II,IK)
0179          300   CONTINUE
0180                IF(IK.EQ.0) GO TO 2
0181                IIK=IK-1
0182                CALL VMULFF(TDM,DM,IIK,NR,IIK,NV,NR,TXX,NV,IER)
```

```
0183                NZ=IK*(IIK+2)/2
0184                DO 304 K=1,NZ
0185          304   YA(K)=0.
0186                K=1
0187                DO 305 I=1,IIK
0188                DO 305 J=1,I
0189                YA(K)=TXX(I,J)
0190          305   K=K+1
0191                IY=6
0192                CALL LEQT1P(YA,1,IIK,B1,NV,IY,D1,D2,IER)
0193                TOTL=0.
0194                P=IIK
0195                QIR=QIR+P
0196                RIR=RIR+P*P
0197                AB=0.
0198                DO 350 I=1,IIK
0199          350   AB=AB+B1(I)
0200                BAB=AB/P
0201                DO 351 I=1,IIK
0202          351   TOTL=TOTL+(B1(I)-BAB)**2
0203                ZNR=NR
0204                TOTL=TOTL/(ZNR-P)
0205                BESQ=0.
0206                DO 320 I=1,IIK
0207          320   BESQ=BESQ+B1(I)**2
0208                ZKK=P*TOTL/BESQ
0209                ZAP=.15
0210                IF(ZKK.LT.ZAP) GO TO 321
0211                ZKK=ZAP
0212          321   CALL VMULFF(TDM,DM,IIK,NR,IIK,NV,NR,TXX,NV,IER)
0213                DO 330 I=1,IIK
0214          330   TXX(I,I)=TXX(I,I)+ZKK
0215                K=1
0216                DO 340 I=1,IIK
0217                DO 340 J=1,I
0218                VA(K)=TXX(I,J)
0219          340   K=K+1
0220                IY=6
0221                CALL LEQT1P(VA,1,IIK,RB,NV,IY,D1,D2,IER)
          CALCULATE RSS FOR RIDGE
```

```
0222                RSSR=0.0
0223                DO 360 I=1,NR
0224                YHAT=0.0
0225                DO 370 J=1,IIK
0226          370   YHAT=YHAT+DM(I,J)*RB(J)
0227          360   RSSR=RSSR+(RVEC(I,NV)-YHAT)**2
0228                K=1
0229                DO 383 I=1,NI
0230                IF(ICRV(I).EQ.0) GO TO 382
0231                REB(I)=RB(K)
0232          381   K=K+1
0233                GO TO 383
0234          382   REB(I)=0.0
0235          383   CONTINUE
0236                CALL MESS(BETA,REB,RSSAL,RSSR,IIK,NI,N1,N2,SIGM,R,RMSE,RCOST,
                   1RCVC,RMSC,RCOS,NR,RAA)
0237                ARSSR=ARSSR+RSSR
0238                ARMSE=ARMSE+RMSE
0239                ARCOST=ARCOST+RCOST
0240                ARCVC=ARCVC+RCVC
0241                ARMSC=ARMSC+RMSC
0242                RACOS=RACOS+RCOS
                C ************************************
                C BAYES REGRESSION                    *
                C ************************************
0243                SRS=SQRT(RSSAL)
0244                CIA=RSSAL/3.
0245                ZK=NI
0246                K=NI
0247                IV=1000000.
0248                IV3=1000000.
0249                IND(1)=0
0250                M=1
0251          401    M=M+1
0252                IND(M)=IND(M-1)+1
0253          402   I1=IND(M-1)+1
0254                I2=IND(M)+1
0255                I3=IND(M)
0256                KP=K+1
0257                M1=M-1
0258                IQ=2
0259                CALL SEMI(I1,I2,I3,C,KP,IND,IQ,M1)
```

```
0260                RSSCA=C(I2,NV,NV)
0261                ZP=M-1
0262                TV1=RSSCA+(ZP/ZK)*SRS
0263                IF(TV1-TV)    61,62,62
0264          61    DO 610 I=1,NI
0265          610   EB(I)=0.0
0266                DO 620 L=2,M
0267                J=IND(L)
0268           620   EB(J)=C(I2,J,KP)
0269                IKP=ZP
0270                BRSS=RSSCA
0271                TV=TV1
0272          62    TV2=RSSCA+(ZP/ZK)*CIA
0273                  IF(TV2-TV3)   619,629,629
0274          619   DO 6109 I=1,NI
0275          6109  EE(I)=0.0
0276                DO 6209 L=2,M
0277                J=IND(L)
0278           6209  EE(J)=C(I2,J,KP)
0279                JKP=ZP
0280                GRSS=RSSCA
0281                TV3=TV2
0282          629   IF(IND(M).LT.K) GO TO 401
0283                M=M-1
0284                IND(M)=IND(M)+1
0285                IF(M.GT.1) GO TO 402
0286                CALL MESS(BETA,EB,RSSAL,BRSS,IKP,NI,N1,N2,SIGM,R,BMSE,BCOST,
                   1BCVC,BMSC,BCOS,NR,RAA)
0287                ABRSS=ABRSS+BRSS
0288                ABMSE=ABMSE+BMSE
0289                ABCOST=ABCOST+BCOST
0290                ABCVC=ABCVC+BCVC
0291                ABMSC=ABMSC+BMSC
0292                PIKP=IKP
0293                QIB=QIB+PIKP
0294                CIB=CIB+PIKP*PIKP
0295                CALL MESS(BETA,EE,RSSAL,GRSS,JKP,NI,N1,N2,SIGM,R,BMS2,BCOST,
                   1BCV2,BMS2,BCOS2,NR,RAA)
0296                ABRSY=ABRSY+GRSS
0297                ABMSY=ABMSY+BMS2
0298                ABCOSY=ABCOSY+BCOS2
0299                ABCVY=ABCVY+BCV2
0300                ABMSY=ABMSY+BMS2
```

```
0301                PKP=JKP
0302                QIY=QIY+PKP
0303                AIB=AIB+PKP*PKP
0304                NOR=NOR+1.
0305                IF(NOR.GE.NUN) GO TO 900
0306                IF(NOR.GE.1) GO TO 2
0307          900   ZNOR=NOR
0308                AVRSSF=ARSSF/ZNOR
0309                AVMSE=AMSE/ZNOR
0310                AVCOST=ACOST/ZNOR
0311                AVCVC=ACVC/ZNOR/R
0312                AVMSC=AMSC/ZNOR
0313                AVMSC=SIGM(N2)*(1-R)/AVMSC
0314                AVRSSR=ARSSR/ZNOR
0315                AVRMSE=ARMSE/ZNOR
0316                AVRCOS=ARCOST/ZNOR
0317                AVRCVC=ARCVC/ZNOR/R
0318                AVRMSC=ARMSC/ZNOR
0319                AMSCR=SIGM(N2)*(1-R)/AVRMSC
0320                AVBRSS=ABRSS/ZNOR
0321                AVBMSE=ABMSE/ZNOR
0322                AVBCOS=ABCOST/ZNOR
0323                AVBCVC=ABCVC/ZNOR/R
0324                AVBMSC=ABMSC/ZNOR
0325                AMSCB=SIGM(N2)*(1-R)/AVBMSC
0326                YVBRSS=ABRSY/ZNOR
0327                YVBMSE=ABMSY/ZNOR
0328                YVBCOS=ABCOSY/ZNOR
0329                AVBCVY=ABCVY/ZNOR/R
0330                YVBMSC=ABMSY/ZNOR
0331                YMSCB=SIGM(N2)*(1-R)/YVBMSC
0332                FAC=FACOS/ZNOR
0333                RAC=RACOS/ZNOR
0334                BAC=BACOS/ZNOR
0335                FVAR=QIF/ZNOR
0336                RVAR=QIR/ZNOR
0337                BVAR=QIB/ZNOR
0338                BVAR2=QIY/ZNOR
0339                WRITE(6,866) FVAR,RVAR,BVAR,BVAR2
0340                SFIF=FIF/ZNOR
```

```
0341            SRIR=RIR/ZNOR
0342            SCIB=CIB/ZNOR
0343            SAIB=AIB/ZNOR
0344            WRITE(6,866)  SFIF,SRIR,SCIB,SAIB
0345      866   FORMAT(//10X,'P FORW   RIDGE   BAYES1   BAYES2',/12X,F5.1,2X,
               1F5.1,5X,F5.1,6X,F5.1//)
0346            WRITE(6,2100) AVRSSF,AVMSE,AVCOST,AVCVC,AVMSC,AVRSSR,AVRMSE,
               1AVRCOS,AVRCVC,AMSCR,AVBRSS,AVBMSE,AVBCOS,AVBCVC,AMSCB
0347      2100  FORMAT(20X,'    RSS       MSE      COST       I1     ',
               1'     I2          '/14X,'FWD',3X,5(F9.3,1X)    /14X,'RIDG',2X,
               25(F9.3,1X)/14X,'BAYES',1X,5(F9.3,1X))
0348            WRITE(6,2101) AVRSSF,AVMSE,FAC,AVCVC,AVMSC,AVRSSR,AVRMSE,RAC,
               1AVRCVC,AMSCR,YVBRSS,YVBMSE,YVBCOS,AVBCVY,YMSCB
0349      2101  FORMAT(20X,'    RSS       MSE      COST2      I1     ',
               1'     I2          '/14X,'FWD',3X,5(F9.3,1X)    /14X,'RIDG',2X,
               25(F9.3,1X)/14X,'BAYES',1X,5(F9.3,1X))

0350      5020  CONTINUE
0351            STOP
0352            END
```

SEMI

```
0001            SUBROUTINE SEMI(IB,IS,IP,C,KP,IND,IA,IZ)
0002            DIMENSION C(11,11,11),IND(11)
0003            C(IS,IP,IP)=1./C(IB,IP,IP)
0004            CALL GAUSS(IB,IS,IP,C,KP)
0005            IF (IA.GT.IZ) GO TO 2
0006            LB=IP+1
0007            DO 1 L=IA,IZ
0008            B=C(IB,IND(L),IP)/C(IB,IP,IP)
0009            C(IS,IND(L),IND(L))=C(IB,IND(L),IND(L))+B*C(IB,IND(L),IP)
0010            DO 1 M=LB,KP
0011      1     C(IS,IND(L),M)=C(IB,IND(L),M)-B*C(IB,IP,M)
0012      2     RETURN
0013            END
```

GAUSS

```
0001            SUBROUTINE GAUSS(IB,IS,IP,C,KP)
0002            DIMENSION C(11,11,11)
0003            LB=IP+1
0004            DO 1 L=LB,KP
0005            C(IS,IP,L)=C(IB,IP,L)/C(IB,IP,IP)
0006            DO 1 M=L,KP
0007      1      C(IS,L,M)=C(IB,L,M)-C(IB,IP,M)*C(IS,IP,L)
0008            RETURN
0009            END
```

MESS

```
0001                  SUBROUTINE MESS(BETA,EBETA,RSS1,RSS2,KP,NI,N1,N2,SIGM,RR,ZMSE,
                     1COST,CVC,ZMSC,COS2,NR,RAA)
0002                  DIMENSION BETA(NI),EBETA(NI),SIGM(N2),RAA(11,11)
0003                  NV=11
0004                  III=0
0005                  QDR1=0.0
0006                  QDR2=0.0
0007                  QDR3=0.0
0008                  QDR4=0.0
0009                  V1=0.0
0010                  V2=0.
0011                  ZMSE=0.0
0012                  DO 10 I=1,NI
0013                  DIFF=BETA(I)-EBETA(I)
0014                  ZMSE=ZMSE+DIFF**2
0015                  KK=I-1
0016                  IJK=I+N1
0017                  V1=V1+EBETA(I)*SIGM(IJK)
0018                  IF(KK.EQ.0) GO TO 20
0019                  DO 11 J=1,KK
0020                  II=KK*I/2+J
0021                  DIFF1=BETA(J)-EBETA(J)
0022                  QDR1=QDR1+EBETA(I)*EBETA(J)*SIGM(II)
0023            11    QDR3=QDR3+DIFF*DIFF1*SIGM(II)
0024            20    III=III+I
0025                  QDR2=QDR2+EBETA(I)**2*SIGM(III)
0026            10    QDR4=QDR4+DIFF**2*SIGM(III)
0027                  DO 25 I=1,NI
0028            25    V2=V2+EBETA(I)*RAA(NV,I)
0029                  QDR=2.*QDR1+QDR2
0030                  QDRR=2.*QDR3+QDR4
0031                  DENO=QDR*SIGM(N2)
0032                  IF(DENO.NE.0.) GO TO 15
0033                  DENO=.001
0034            15    V1=V1**2
0035                  CVC=V1/DENO
0036                  ZMSC=SIGM(N2)*(1.-RR)+QDRR
0037                  ZP=KP
0038                  ZK=NI
0039                  COST=RSS2+SQRT(RSS1)*ZP/ZK
0040                  COS2=RSS2+RSS1*ZP/(3.*ZK)
0041                  RSS2=1-(1-V2)*(NR-1)/(NR-ZP)
0042                  RETURN
0043                  END
```

Appendix F

Eigenvalues of the three interpredictor correlation matrices

For the low interpredictor correlations:

$\lambda_1 = 2.77$ $\lambda_2 = 1.03$ $\lambda_3 = .95$ $\lambda_4 = .91$
$\lambda_5 = .89$ $\lambda_6 = .80$ $\lambda_7 = .77$ $\lambda_8 = .68$
$\lambda_7 = .62$ $\lambda_{10} = .59$

Mean of the eigenvalues = 1.00

Variance of the eigenvalues = .37

For the medium interpredictor correlations:

$\lambda_1 = 5.47$ $\lambda_2 = .73$ $\lambda_3 = .65$ $\lambda_4 = .61$
$\lambda_5 = .59$ $\lambda_6 = .50$ $\lambda_7 = .47$ $\lambda_8 = .38$
$\lambda_9 = .32$ $\lambda_{10} = .29$

Mean = 1.00

Variance = 2.24

For the high interpredictor correlations:

$\lambda_1 = 7.25$ $\lambda_2 = .54$ $\lambda_3 = .49$ $\lambda_4 = .40$
$\lambda_5 = .38$ $\lambda_6 = .27$ $\lambda_7 = .25$ $\lambda_8 = .18$
$\lambda_9 = .13$ $\lambda_{10} = .11$

Mean = 1.00

Variance = 4.35

PLEASE NOTE:

Page 122 is lacking in number only. No text is missing. As received from the Graduate School.

UNIVERSITY MICROFILMS INTERNATIONAL

REFERENCES

Aitkin, M.A., Simultaneous Inference and the Choice of Variable Subsets in Multiple Regression, Technometrics, 16, no. 2, 1974, 221-227

Allen, D.M., Mean Square Error of Prediction as a Criterion for Selecting Variables, Technometrics, 13, 1971, 469-475

Ando, A., and Kaufman, G.M., Bayesian Analysis of the Independent Multinormal Process, Journal of American Statistical Assoc., 60, 1965, 347-358

Bartlett, M.S., On the Theory of Statistical Regression, Proceedings of the Royal Society of Edinburgh, 1933, 53, 260-283

Beale, E.M.L., Note on Procedure for Variable Selection in Multiple Regression, Technometrics, 12, 1970, 909-914

Beale, E.M.L., Kendall, M.D., and Mann, D.W., The Discarding of Variables in Multivariate Analysis, Biometrics, 54, 357-366

Browne, M.W., Comparison of Single Sample and Cross-Validation Methods for Estimating the Mean Squared Error of Prediction in Multiple Linear Regression, British Journal of Mathematical Statistical Psychology, vol. 28, 1, 1975, 112-120

Browne, M.W., Precision of Prediction, Research Bulletin 69-69, Princeton: Educational Testing Service, 1969.

Burkett, G.R., A Study of Reduced Rank Model for Multiple Prediction, Psychometric Monographs, no 12, 1964

Brownlee, K.A., Statistical Theory and Methodology in Science and Engineering, New York: Wiley & Sons, 1965

Coniffe, D. and Stone, J., A Critical Review of Ridge Regression, Statistician, 22, 2, 1974, 181-187

Darlington, R.B., Multiple Regression in Psychological Research and Practice, Psychology Bulletin, 1968, 3, 161-181

Davies, ed., Statistical Methods in Research and Production, New York: Hafner, 1958

DeGroot, M.H., Optimal Statistical Decisions, New York: McGraw-Hill, 1970

Dempter, A.P., Model Searching and Estimation in the Logic of Inference, in Foundations of Statistical Inference, edited by Godambe, V.P. and Spratt, D.A., Toronto: Holt, Rinehart & Winston, 1971

Draper, N. and Smith, H., Applied Regression Analysis, New York: Wiley & Sons, 1966

Dwyer, P.S., The Square-Root Method and its Use in Correlation and Regression, J.A.S.A., 1945, 40, 493-503

Effromyson, M.A., Multiple Regression Analysis, in Ralston & Wilf, Mathematical Methods for Digital Computers, New York: Wiley & Sons, 1960

Furnival, G.M., and Wilson, R.W., Regression by Leaps and Bounds, Technometrics, vol. 16, no. 4, 1974, 499-511

Garside, M.J., The Best Subset in Multiple Regression Analysis, Applied Statistical Journal of the Royal Statistical Society, 14, 1965, 196-200

Gorman, J.W., and Toman, R.J., Selection of Variables for Fitting Equations, Technometrics, 8, 1966, 27-52

Graybill, F.A., Introduction to Linear Statistical Models, vol. 1, New York: McGraw-Hill, 1961

Guilkey, D.K., and Murphy, J.L., Directed Ridge Regression Techniques in Cases of Multicollinearity, J.A.S.A., 70, December 1975, 769-775

Hardy, G.H., Littlewood, J.E., and Polya, G., Inequalities, Cambridge: Cambridge University Press, 1934

Herzburg, P.A., The Parameters of Cross-Validation, Psychometric Mono. Supplement, no. 16, 1969, 34

Hocking, R.R., Criteria for the Selection of a Subset Regression: Which One Should be Used?, Technometrics, 15, 967-970

Hocking, R.R., and Leslie, R.N., Selection of the Best Subset in Regression Analysis, Technometrics, 9, 1967, 531-540

Hoerl, A.E., and Kennard, R.W., Ridge Regression, Technometrics, 12, 1970, 55-67, 69-82

Horst, P., and MacEwan, C., Predictor Elimination Techniques for Determining Multiple Predictive Batteries, Psychometric Reports, 7, 1960, 19-50

Jeffreys, H., Theory of Probability, Oxford: Clarendon Press, 1961

Lindley, D.V., The Choice of Variables in Multiple Regression, Journal of the Royal Statistical Society B, 30, no. 11, 1968, 31-66

Lindley, D.V., Bayesian Statistics: Introduction to Probability and Statistics from a Bayesian Point of View, Cambridge: Cambridge University Press, 1965

Lindley, D.V., Making Decisions, London: Wiley & Sons, 1971

Lindley, D.V., and Smith, A.F.M., Bayes Estimates for the Linear Model, Journal of the Royal Statistical Society, ser. B, 34, no. 1, 1972, 1-18

Lord, F.M., Efficiency of Prediction When a Progression Equation From One Sample is Used in a New Sample, Princeton: Educational Testing Service, 1950 (cited in Darlington)

Lord, F.M., and Novick, M.R., Statistical Theories of Mental Test Scores, New York: Addison Wesley, 1968

Mallows, C.L., Choosing Variables in Linear Regression, presented at The Institute of Mathematical Statistics, Kansas, 1964

Mallows, C.W., Some Comments on Cp, Technometrics, 15, no. 4, 1973, 661-675

Mantel, N., Why Stepdown Procedures in Variable Selection, Technometrics, 12, 1970, 621-625

Marquadt, D.W., Generalized Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation, Technometrics, 12, August 1970, 591-612

Marquadt, D.W. and Snee, R.D., Ridge Regression in Practice, American Statistician, 1975, 1, 3-20

Mayer, L.S., and Willke, T.A., On Biased Estimation in Linear Models, Technometrics, 15, August 1973, 497-508

McDonald, G.C., and Galarneau, D.I., A Monte Carlo Evaluation of Some Ridge-Type Estimators, J.A.S.A., June 1975, 407-416

Morgan, J.A. and Tatar, J.I., Calculation of the Residual Sum of Squares for All Possible Regressions, Technometrics, 14, 317-325

Raiffa, H., and Schlaifer, R., Applied Statistical Decision Theory, Boston: Harvard Business School, 1961

Rock, D.A., Linn, R.L., Evans, F.R., and Patrick, C.A., A Comparison of Predictor Selection Techniques Using Monte Carlo Methods, Educational Psychological Measurement, Winter 1970, 873-884

Sclove, S.L., Improved Estimators for Coefficients in Linear Regression, J.A.S.A., 63, 1968, 596-606

Spjotvall, E., Multiple Comparisons of Regression Functions, Ann. Math. Stat., 43, 1076-1088

Summerfield, A., and Lubin, A., A Square-Root Method of Selecting a Minimum Set of Variables in Multiple Regression, Psychometrika, 1951 15, 271-284

Theil, H., Economic Forecasts and Policy, second edition, Amsterdam: N. Holland Publishing Co., 1961

Theil, H., On the Use of Incomplete Prior Information in Regression Analysis, J.A.S.A., 58, June 1963, 401-414

Theobold, C.M., Generalization of Mean Square Error Applied to Ridge Regression, Journal of the Royal Statistical Society, B, 36, 1974, 103-106

Wermuth, N.E., Empirical Comparison of Regression Methods. Unpublished doctoral dissertation, Harvard University, 1972

Wherry, R.J., A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation, Annals of Mathematical Statistics, 1931, 2, 440-457

Zellner, A., An Introduction to Bayesian Inference in Econometrics, New York: Wiley & Sons, 1971

Zellner, A., and Chetty, V.K., Prediction and Decision Problems in Regression Models from the Bayesian Point of View, J.A.S.A., 1965, 60, 608-619

Zellner, A. and Vandaele, W., Bayes-Stein Estimators for K-Means, Regression and Simultaneous Equation Models (cited in Wermuth, N.E.)