Recommended citation:
Herzberg, P. A. (1969). *The Parameters of Cross-Validation* (Psychometric Monograph No. 16). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN16.pdf

THE PARAMETERS OF CROSS-VALIDATION

Paul A. Herzberg

YORK UNIVERSITY
TORONTO

Printed by The William Byrd Press, Richmond, Va.

PREFACE

This monograph was submitted as a doctoral dissertation at the University of Illinois, Urbana. The author is indebted to Professors L. R Tucker and J. S. Wiggins for their generous help. The work on this project was partially supported by the Office of Naval Research under contracts Nonr 1834(39) and US NAVY/00014-67-A-0305-0003 with the University of Illinois.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Multiple Regression and Cross-Validation

A problem common to many areas of psychology is the prediction of a person's score on one variable from his scores on a number of other variables. The variable that is to be predicted is called the criterion and the other variables are called predictors. Many methods have been developed to combine predictor scores in order to optimize the prediction of the criterion. A common procedure is to obtain a sample of subjects with known predictor and criterion scores (the derivation sample) and to calculate the *linear* combination of the predictor scores that best predicts the criterion scores. By "best" is usually meant "least squared error," which means that the sum (over subjects) of the squared deviations of the observed from the predicted criterion score is a minimum. The optimizing coefficients of the predictor scores are called the multiple regression weights and are calculated from the normal equations which express the minimization conditions [Anderson, 1958; Kendall & Stuart, 1961].

When multiple regression is used to compute predictor weights, a multiple correlation may be calculated. The multiple correlation is the Pearson product-moment correlation, in the sample, between the optimal linear combination of the predictors and the criterion variable. The multiple correlation is thus a measure of the degree of relationship between the predictors and the criterion. However the multiple correlation is a biased estimate of this relationship and is generally larger than the true population multiple correlation. The bias occurs because the process of minimizing the average squared error in prediction is equivalent to maximizing the correlation between the linear combination of the predictors and the criterion. Due to the finite size of the sample, the optimizing linear combination will be fitted to the idiosyncracies of the sample and will generally result in a higher multiple correlation than the population multiple correlation.

One problem in the application of multiple correlation techniques is therefore the estimation of the true multiple correlation from the biased sample multiple correlation. In the next section it will be shown that there are two population correlations which must be distinguished. A number of formulas for correcting the sample multiple correlation is known. However these formulas require assumptions which are often difficult to satisfy and, therefore, many early investigators estimated the population correlation by applying to a second sample the regression weights calculated in an original

sample. They found that the correlation between the regression function and the criterion in the second sample was less than the original sample multiple correlation. This technique became known as cross-validation of the predictor weights or simply as cross-validation [Mosier, 1951]. The correlation in the second sample is called the cross-validity. The first sample is known as the *derivation* sample; the second is the *validation* sample. An obvious addition to the cross-validation method is to repeat the calculations, interchanging the roles of the first and second sample. We shall call this technique double cross-validation.

This study was designed to investigate:

(a) the accuracy of the cross-validity as an estimate of the population correlation,
(b) the effectiveness of two reduced rank methods for estimating predictor weights, and
(c) the effect of the variation of some parameters of the population distribution on the results of (a) and (b).

The estimation of the population correlation is described in more detail in Section 1.2. The reduced rank methods are introduced in Section 1.3. Finally, the study of the effect of variation of population parameters by a simulation technique is introduced in Section 1.4.

1.2 Estimates of Validity

Let the predictor variables be x_1, x_2, $\cdots$, x_n and let the criterion variable be y. Then the regression function in the population is

$$\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \beta_0 . \tag{1.2.1}$$

The constant term in the equation is β_0; the β_i are called the regression weights. Two models for the predictors are possible, the regression model and the correlation model [Ezekiel & Fox, 1959, pp. 279-281]. In the regression model, the values of the predictor variables are fixed and only the criterion is a random variable. A more realistic model for most multivariate work in psychology is to assume that both the predictors and the criterion are random variables (the correlation model). Under the null hypothesis of zero multiple correlation the distributional theory is identical for the two models. However when the null hypothesis is not true the distributions are different under the two models. Since the distributional theory is much more complicated under the correlation model, most investigators in psychology [e.g. Burket, 1964] have continued to use the regression model hoping that there will be little practical difference between the two models.

Regression equations can also differ in whether the constant term, β_0, is included. In the case of the regression model, the constant term is really indistinguishable from the other terms in the equation since a predictor

variable, x_0, may be defined as the constant 1.0. Then the constant term may be written as $\beta_0 x_0$. Therefore, formulas developed for the constant = 0 case

$$\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{1.2.2}$$

may be modified for the constant $\neq 0$ case (1.2.1) by simply replacing n by $n + 1$.

In the correlation model this simple correspondence between the zero and non-zero constant cases does not hold since $x_1, \cdots, x_n$ are random variables while x_0 is fixed. Inclusion of a constant term does not affect the multiple correlation or the correlation between the regression function and any other variable. Thus in studies such as the present one which emphasize correlation measures, it is simplest to set the constant term to zero. However if the mean square error of prediction is used as a measure of accuracy of prediction, it is very important to state whether the constant term is included in the regression.

Let ρ be the population multiple correlation, σ_y the population standard deviation of y, and $\sigma_{(y-\hat{y})}$ the population standard deviation of the error in prediction $(y - \hat{y})$ where $\hat{y}$ is the regression function (1.2.1) with weights calculated from the normal equations. Then

$$\rho^2 = 1 - \frac{\sigma^2_{(y-\hat{y})}}{\sigma^2_y}. \tag{1.2.3}$$

A similar equation holds in the sample, relating the squared sample correlation, r^2, to the mean squared error of prediction, MSE, and the standard deviation of the sample, s_y:

$$r^2 = 1 - \frac{\text{MSE}}{s^2_y}. \tag{1.2.4}$$

The first estimation of ρ^2 in the psychological literature [Larson, 1931] is by the following formula:

$$\text{Est}\,(\rho^2) = 1 - \frac{N}{N - n}(1 - r^2) \tag{1.2.5}$$

where N is the sample size. Larson does not give a derivation of this formula but Wherry [1931] showed that it follows (in the regression model) from estimating σ^2_y by s^2_y and estimating $\sigma^2_{(y-\hat{y})}$ by $(\text{MSE})N/(N - n)$. The substitution of these two estimates into (1.2.3) and the use of (1.2.4) gives (1.2.5). In order to improve this estimate, Wherry estimated σ^2_y by $s^2_y N/(N - 1)$ rather than by s^2_y. The resulting formula is

$$\text{Est}\,(\rho^2) = 1 - \frac{N - 1}{N - n}(1 - r^2). \tag{1.2.6}$$

Larson and Wherry compared their estimates with cross-validities and Wherry showed that (1.2.6) is superior to (1.2.5).

It is not entirely clear how Larson and Wherry handled the constant term in the regression function. Formula (1.2.6) is strictly applicable to a zero constant term. When the constant term is not zero, the unbiased estimate of $\sigma^2_{(y-\hat{y})}$ is (MSE)N/(N − n − 1) so that the estimate of ρ^2 is

$$\text{Est}(\rho^2) = 1 - \frac{N-1}{N-n-1}(1 - r^2). \tag{1.2.7}$$

Formula (1.2.7) is often referred to as Wherry's formula even though his original formula was (1.2.6). Formula (1.2.7) is not an unbiased estimate of ρ^2 since the ratio of two unbiased estimates is not unbiased. However, unbiased estimates of ρ^2 are not always desirable, for, if the true $\rho^2 = 0$, an unbiased estimate must take on both negative and positive values even though a multiple correlation is always positive.

The multiple correlation, ρ, is the correlation, in the population, of the criterion and the regression function calculated in the population. In applications, the population regression function can never be known and one is more interested in how effective the *sample* regression function is in *other* samples. A measure of this effectiveness is r_c, the sample cross-validity. For any given regression function, r_c will vary from validation sample to validation sample. The average value of r_c will be approximately equal to the correlation, in the *population*, of the sample regression function with the criterion. This correlation is the population cross-validity, ρ_c. Wherry's formula estimates ρ rather than ρ_c. Lord [1950] and Nicholson [1960] derived an unbiased estimate of the population mean square error of a sample regression function. Using this estimate of MSE, an estimate of ρ_c^2 is

$$\text{Est}(\rho_c^2) = 1 - \frac{N-1}{N-n-1}\frac{N+n+1}{N}(1 - r^2). \tag{1.2.8}$$

This formula applies to the *regression* model with a constant term. Darlington [1968] modified this formula for the *correlation* model with a constant term. His formula is

$$\text{Est}(\rho_c^2) = 1 - \frac{N-1}{N-n-1}\frac{N-2}{N-n-2}\frac{N+1}{N}(1 - r^2). \tag{1.2.9}$$

This formula is based on the assumption that the predictors and criterion have a multivariate normal distribution.

It would be possible to derive a similar formula for $\text{Est}(\rho_c^2)$ in the multivariate normal case with no constant term. It is not, however, the purpose of this study to derive the best estimate of ρ_c^2 for it is not clear what properties such an estimator should have, particularly since an unbiased estimator

has the defects mentioned above. It is more interesting to study the accuracy of the cross-validity as an estimate of ρ_c and ρ.

Returning to estimating ρ, Wishart [1931] calculated the moments of the distribution of r^2 for the multivariate normal distribution. The expected value of r^2 is

$$(1.2.10) \qquad E(r^2) = 1 - \frac{N-n-1}{N-1}(1-\rho^2)F(1, 1, (N+1)/2, \rho^2)$$

where F(a, b, c, x) is the hypergeometric function. Using the first two terms of the expansion of this function, (1.2.10) reduces to

$$(1.2.11) \qquad E(r^2) = 1 - \frac{N-n-1}{N-1}(1-\rho^2) - \frac{N-n-1}{N-1}\frac{2}{N+1}\rho^2(1-\rho^2).$$

Olkin and Pratt [1958] showed that an unbiased estimate of ρ^2 is

$$(1.2.12) \qquad \text{Est}(\rho^2) = 1 - \frac{N-3}{N-n-1}(1-r^2)F(1, 1, (N-n+1)/2, 1-r^2),$$

which, neglecting terms in $1/N^2$, is

(1.2.13)

$$\text{Est}(\rho^2) = 1 - \frac{N-3}{N-n-1}(1-r^2) - \frac{N-3}{N-n-1}\frac{2}{N-n+1}(1-r^2)^2.$$

The Wherry estimate (1.2.7) is almost identical to the first two terms of this series.

Darlington [1968] has carefully distinguished the four correlations ρ, ρ_c, r, and r_c. The smallest of these, ρ_c and r_c, are the validity of the sample regression function in the population and another sample, respectively. The average, over many samples, of the cross-validity, r_c, will be approximately equal to ρ_c. The next smallest correlation is ρ, the population multiple correlation or the validity of the population regression function in the population. On the average, the largest correlation is the sample multiple correlation r, which is the validity of the sample regression function in the derivation sample. The relationships may be summarized as follows:

$$(1.2.14) \qquad E(r_c) \simeq \rho_c < \rho < E(r).$$

Empirical confirmation of (1.2.14) is presented in Section 3.3.

1.3 Improvement of Prediction

It is well known that adding predictors to a regression equation increases both the sample and population multiple correlations. However, the greater the number of predictors, n, the more unstable are the sample regression weights and the lower are the sample cross-validity, r_c, and the

population cross-validity, ρ_c. The decrease in estimated ρ_c follows from (1.2.9).

A second difficulty with a large number of predictors in multiple regression is that a subset of them would probably do just as well, if the subset could be determined. For predictions in applied psychology, e.g. personnel selection, it is undesirable to have to make a large number of measurements on each individual in order to make accurate predictions. Furthermore, the weights for a subset of predictors would be more stable in future samples due to the smaller n.

There are several ways to select a subset of predictors. The best selection procedure is stepwise regression in which predictors are added to the regression, one at a time, until there is no significant additional prediction. Other selection procedures are shown by Darlington [1968] to be inferior to the stepwise method.

Another way to reduce the number of predictors in the regression function is to use a few linear combinations of the predictors rather than the predictors themselves. Two such methods, called reduced rank methods, are considered in this study. In the first method [Horst, 1941], the largest principal components of the predictors [Anderson, 1958] are entered in the regression function. Since the principal components may be expressed as linear combinations of the predictors, the regression function may be transformed to a linear combination of the predictors. Hence the full set of predictor variables is used but only through the intermediary of a few principal components. These components may be interpreted psychologically, and it may be possible to select predictors loading highly on the components as a subset to use in future prediction. In this way reduced rank prediction can lead to a reduction of the size of the predictor battery.

In his 1941 paper, Horst also suggested that the predictors could be represented as a linear function of common and unique factors rather than as a linear function of the principal components. This factor analytic model is more difficult to treat because of the difficulty of estimating the factors as linear combinations of the predictors. Unlike Horst's first method, factor analysis is not a reduced rank method. The factor analytic model for regression calculations was studied by Leiman [1951] with some success but will not be considered further in this study.

Before outlining the second reduced rank procedure, let us consider a study by Burket [1964] comparing a number of regression methods in a large data sample. He compared two stepwise selection procedures [Efroymson, 1960; Horst & MacEwan, 1960], the largest principal components method, the smallest principal components method [Guttman, 1958] and the criterion-related principal components method [Hotelling, 1957; Massy, 1965]. Guttman proposed the use of the smallest principal components since the solution for the multiple regression weights depends on the inverse of the predictor

intercorrelation matrix, and the largest components of the inverse are the smallest components of the original matrix. Hotelling and Massy suggested that the principal components which are entered into the regression function should be those components correlating maximally with the criterion rather than those of largest variance. Burket compared these five methods for several criteria and in several subsamples of his total sample. He found that the largest principal components method was consistently superior to the other four methods. One purpose of the present study is to show under what conditions this superiority can be expected to hold.

The second reduced rank method, prediction from the principal predictors, was developed from the following considerations. The principal components of the predictors may not be highly related to the criterion since the components are determined solely from the intercorrelations of the predictors. It would be desirable to find linear combinations of the predictors which are strongly related to the criterion. The Hotelling and Massy method employed by Burket finds these linear combinations by computing the correlation of each principal component with the criterion and entering into the multiple regression only those components with the highest correlations. However a more effective procedure might be to find those linear combinations of the predictors (not necessarily the principal components of the predictors) which are maximally correlated with the criterion.

In the single criterion case, this problem is trivial since there is only one linear combination of the predictors maximally correlated with the criterion and all other orthogonal combinations are uncorrelated with the criterion. This combination is simply the regression function, i.e. the predicted criterion, using multiple regression on all the predictors. Therefore, in the single criterion case, nothing new is found by considering linear combinations of the predictors maximally correlated with the criterion.

Consider, however, prediction of several criteria from a common set of predictors. Examples of such multiple criteria are the prediction of success in several academic curricula by using a battery of aptitude tests or the prediction of a number of social criteria using scales from a personality test [Hase & Goldberg, 1967]. In particular, let us suppose that we wish to predict each of the criteria equally well. Then Tucker [1957] has developed a method which discovers those linear combinations of the predictors maximally related to the set of criteria. These combinations are called the *principal predictors*. The largest of the principal predictors may be entered into the regression equations for each criterion. The principal predictors have the property that, for a fixed number of linear combinations entered into each regression, the average squared multiple correlation is greater for the principal predictors than for any other linear combinations entered into the regression.

The principal predictors were developed by Tucker as a convenient way

to summarize a large number of predictor scores by a few criterion-related predictor scores. The principal predictors also provide a useful conceptualization of the relationship of a set of predictors to a set of criteria. In the present study, on the other hand, the principal predictors are compared with the principal components as reduced rank prediction methods.

In any prediction calculation, each criterion variable may be divided into two parts—one part is predictable from the set of predictors and the other part is unpredictable from these predictors. When there are several criteria the predictable parts of the criteria are themselves a set of variables which have principal components. These principal components are the principal predictors. It is important not to confuse the principal components of the predictors, previously discussed, with the principal components of predictable parts of the criteria, which are called the principal predictors. The largest principal predictor accounts for the largest portion of the predictable variation in the criteria. The next largest principal predictor accounts for the next largest portion, and so on. Therefore a few of the principal predictors account for most of the predictable variation in the criteria.

The largest principal predictors may be used as predictors themselves. Then the principal predictors, like the principal components of the predictors, can be expressed in terms of the original predictors. The weights for the original predictors can therefore be calculated. Also, the principal predictors may be interpreted psychologically, which may lead to greater understanding of the relationship between the predictors and the criteria. The principal predictors, unlike the principal components of the predictors, are criterion-related so that variation in the predictors which is unrelated to the criteria is not represented in the principal predictors. In some cases, the predictor variation which is unrelated to the criteria could be large enough to dominate the principal components of the predictors. But this variation is not useful for prediction. This is the reason that prediction from the largest principal predictors may be superior to prediction from the largest principal components.

The two methods, prediction from the principal components of the predictors and from the principal predictors, are called reduced rank methods since, in both cases, a correlation matrix may be approximated by a matrix of lower rank using the largest principal components or the largest principal predictors. In the first method, the correlation matrix of the predictors is approximated, while in the second method, the correlation matrix of the predictable parts of the criteria is approximated. These statements are made more precise in Chapter 2.

Of the prediction methods discussed in this section only three will be considered further in this study:

(a) prediction from the full set of predictors,

(b) prediction from the largest principal components of the predictors, and
(c) prediction from the largest principal predictors.

The last method is possible only when there are several criteria. The first two methods may be used for one or several criteria.

1.4 The Comparison of Prediction Methods

In order to evaluate and compare the prediction methods described in the preceding section it would be desirable to employ mathematical techniques. However the problems are so complex that multivariate statistical theory is unable to solve most of them.

Another approach to these problems has been to apply the different prediction methods to a common body of data and to compare the results [Burket, 1964; Leiman, 1951]. There are definite advantages to this approach. Any conclusions are based on real data and do not depend on the assumptions in a theoretical development being valid. However, there is a major drawback to such empirical techniques. If two or more studies, using different data, disagree in their conclusions, it is difficult to determine what properties of the data sets differ enough between the studies to produce the varied conclusions. Similarly, it is difficult to evaluate the generality of conclusions found in a single study using one set of data.

It is therefore desirable to compare the prediction methods on a wide variety of data sets, differing in a known way in certain parameters. Since it is hard to satisfy this condition with real data, it is proposed that some useful conclusions may be made from the study of artificial or simulated data sets. Such data sets can be readily generated on a computer. The parameters specifying the properties of a data set can be input to the computer and a wide variety of data sets can be generated by varying these input parameters. The prediction method can then be compared in these data. Such a simulation procedure is described in this study.

The simulation experiments consist of four stages of calculations:

Generation of the model. A combined predictor and criterion population covariance matrix is generated subject to certain input parameters. The population model and its parameters are described in Sections 2.2 and 2.3. In the model the predictors and criteria are expressed in terms of the principal predictors.

Generation of two samples. Two samples, each of size N, are obtained from the population generated in the preceding stage. The samples are obtained by generating sample covariance matrices; the method is outlined in Section 2.4. The two samples are used in double cross-validation.

Calculation of predictor weights. The predictor weights are calculated, in each sample, by one or more of the following methods—(a) multiple regression on the predictors, (b) multiple regression on the principal com-

ponents, and (c) multiple regression on the principal predictors. The calculation of these weights is described in Sections 2.5, 2.6, and 2.7, respectively.

Cross-validation of the weights. The weights for each sample (and each method) are cross-validated on the other sample (r_c) and on the population itself (ρ_c). The formulas for the validities are presented in Section 2.8.

CHAPTER 2

THE MATHEMATICAL MODEL AND SAMPLE CALCULATIONS

2.1 Notation

Scalars are denoted by lower case letters (m, ρ). The only exceptions to this convention are N for sample size and the elements of matrices. Scalars may be either numbers or random variables. Column vectors are denoted by lower case italicized letters (x, a). These vectors may be either random variable vectors or vectors of numbers. Row vectors are transposed column vectors, transposition being indicated by priming (x', a'). Matrices are denoted by upper case letters (A, Σ). Transposed matrices are indicated by a prime (A′, Σ'). The (i, j) element of a matrix is denoted by A_{ij} . The identity matrix is I.

The matrix consisting of the first t columns of a matrix A is denoted by $(A)_t$. The first t rows of A are ${}_t(A)$. The vector consisting of the first t elements of a vector b is denoted by ${}_t(b)$.

The population covariance matrix of two random vectors x and y is denoted by Σ_{xy} . The corresponding sample covariance matrix is C_{xy} . When y is known to have only one component, the covariance matrices are column vectors denoted by σ_{xy} and c_{xy} . The variance of a scalar y is denoted by σ_{yy} or c_{yy} . The abbreviation Var( ) is used to denote the variance of the random variable enclosed in parentheses.

The univariate normal distribution with mean = m and variance = v is denoted by N(m, v). The multivariate normal distribution with mean vector = a and covariance matrix = Σ is represented by $N(a, \Sigma)$. Fisher's F distribution, Student's t distribution and the chi distribution are denoted by $F(n_1, n_2)$, $t(n)$, $\chi(n)$, respectively, with the indicated degrees of freedom. The chi distribution is the square root of the chi-squared distribution.

2.2 The General Model for Predictors and Criteria

Let x (n components) be n random variables, called predictors, and let y (m components) be m random variables, called criteria. Let m be less than n as is usually the case in practice. For convenience, normalize all variables x and y to unit variance. Let x and y have a joint multivariate normal distribution with null mean vector and arbitrary covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}. \tag{2.2.1}$$

It can be shown [Herzberg, 1967] that x and y can be written in a special way in terms of (n + m) independent unit variance random variables w. Let w be partitioned as

(2.2.2) $$w' = (w_1'w_2'w_3')$$

where w_1 has m components, w_2 has (n − m) components and w_3 has m components. Then

(2.2.3) $$\begin{matrix} \\ \text{n} \\ \text{m} \end{matrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{matrix} \text{m} & (\text{n}-\text{m}) & \text{m} \\ \end{matrix}\begin{bmatrix} \text{S}_1 & \text{S}_2 & 0 \\ \text{F} & 0 & \text{E} \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}.$$

(The number of rows or columns in the partitioned matrices are appended to the matrix expression above).

Thus the submatrices forming Σ may be written as

(2.2.4) $$\Sigma_{xx} = \text{S}_1\text{S}_1' + \text{S}_2\text{S}_2'$$

(2.2.5) $$\Sigma_{xy} = \Sigma_{yx}' = \text{S}_1\text{F}'$$

(2.2.6) $$\Sigma_{yy} = \text{FF}' + \text{EE}'.$$

This representation of x and y may be understood in the following way. Let y be written as the sum of two terms

(2.2.7) $$y = \hat{y} + e,$$

where $\hat{y}$ is the linear least squares prediction of y from the predictors x. That is,

(2.2.8) $$\hat{y} = B'x = \Sigma_{yx}\Sigma_{xx}^{-1}x.$$

The weight matrix is written as B′ rather than B so that when $m = 1$, $\text{B}' = b'$, the transpose of a column vector. Each component of the predictable part $\hat{y}$ is the best predictor of the corresponding component of y. Now the variables $\hat{y}$ may be transformed to independent unit variance variables w_1, by

(2.2.9) $$\hat{y} = \text{F}w_1$$

where F is orthogonal by columns, so that

(2.2.10) $$\text{F}'\text{F} = \text{D}^2 \text{ (diagonal)}$$

with the diagonal elements of D^2 in descending order. The m variables w_1 are called the *principal predictors* and are, except for normalization, the principal components of $\hat{y}$. The diagonal elements of D^2 are the eigenvalues of $\Sigma_{\hat{y}\hat{y}}$ and are

(2.2.11) $$\text{D}_{kk}^2 = \sum_{j=1}^{m} \text{F}_{jk}^2 .$$

The variables e may also be transformed to independent unit variance variables w_3 by

$$e = Ew_3 . \tag{2.2.12}$$

This latter transformation may be done in many ways. The representation of y is now complete.

The predictors x are also represented as the sum of two terms,

$$\hat{x} = x + d, \tag{2.2.13}$$

where $\hat{x}$ is the part of x linearly predictable from the principal predictors w_1. The relationship of $\hat{x}$ to the principal predictors is

$$\hat{x} = S_1 w_1 . \tag{2.2.14}$$

The residual vector d is related to (n − m) independent unit variance variables w_2 by

$$d = S_2 w_2 . \tag{2.2.15}$$

The fact that the n-vector d is of rank (n − m) is shown in Herzberg [1967].

In summary then, the criteria y are written as the sum of a linear transformation of the principal predictors w_1 and a transformation of m other independent variables w_3. Similarly, the predictors x are written as the sum of a linear transformation of the same principal predictors w_1 and a transformation of (n − m) other independent variables w_2. The association between y and x is expressed through their dependence on the principal predictors w_1. The non-associated parts of x and y are expressed in terms of independent variables w_2 (for x) and w_3 (for y). The total set of variables $w' = (w_1' w_2' w_3')$ are independent unit variance normal.

The matrices S_1 and F are central to the description of the dependence of the criteria on the predictors. Let us first consider F. The squared multiple correlation of the jth criterion y_j with the n predictors is (recall that y_j has unit variance)

$$\rho_j^2 = \mathrm{Var}\,(\hat{y}_j) = (B'\Sigma_{xx}B)_{jj} = (FF')_{jj} = \sum_{k=1}^{m} F_{jk}^2 . \tag{2.2.16}$$

The average, over criteria, of the squared multiple correlation is

$$\begin{aligned} \rho^2 &= (1/m) \sum_{j=1}^{m} \rho_j^2 = (1/m) \sum_{j=1}^{m} \sum_{k=1}^{m} F_{jk}^2 \\ &= (1/m) \sum_{k=1}^{m} D_{kk}^2 \end{aligned} \tag{2.2.17}$$

using (2.2.11). The size or importance of the kth principal predictor for predicting the criteria y can be measured by the sum of squares of the co-

efficients of the kth principal predictor in (2.2.9). The coefficients are the kth column of F. The sum of squares is

$$\sum_{j=1}^{m} F_{jk}^2 = D_{kk}^2 . \tag{2.2.18}$$

Hence the eigenvalues D_{kk}^2 can be interpreted as the total variance of the criteria accounted for by the kth principal predictor. The D_{kk}^2 are, by definition, decreasing with k, so that the first principal predictor accounts for more of the variance of the criteria than any other principal predictor. If, in a certain population, the first eigenvalue D_{11}^2 is very large and the others small, this indicates that most of the prediction of y from x is derived from only one linear combination of the x variables, namely the first principal predictor. On the other hand, if several of the D_{kk}^2 are large, then several independent linear combinations of the predictors are needed in order to get maximum prediction of y from x.

The relation of x to the principal predictors w_1 through the matrix S_1 is quite independent of the matrix F and the quantities D_{kk}^2. Let us define the (n × n) matrix S as the super matrix

$$S = (S_1 S_2). \tag{2.2.19}$$

Thus

$$x = \hat{x} + d = S_1 w_1 + S_2 w_2 = S \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}. \tag{2.2.20}$$

Since the variables w are independent and of unit variance, and the predictors x are of unit variance, the sum of squares of each row of S is 1.0. The sum of squares of the ith row of S_1 is then less than 1.0 and is $\mathrm{Var}(x_i)$, the magnitude of the dependence of the ith predictor on the principal predictors:

$$\mathrm{Var}(\hat{x}_i) = \sum_{k=1}^{m} S_{ik}^2 . \tag{2.2.21}$$

Turning to the columns of S, let $q_k^2 (k = 1, \cdots, m)$ denote the sum of squares of the kth column of S:

$$q_k^2 = \sum_{i=1}^{n} S_{ik}^2 . \tag{2.2.22}$$

q_k^2 is a measure of the total dependence or relation of the predictors x to the kth principal predictor. The q_k^2 are analogous to the D_{kk}^2 since they represent, respectively, the total dependence of the predictors and the criteria on the kth principal predictor. ($\mathrm{Var}(\hat{x}_i)$ and ρ_j^2 are also analogous.) We may average the $\mathrm{Var}(\hat{x}_i)$ in the same way as the ρ_j^2 were averaged in (2.2.17):

$$\pi^2 = (1/n) \sum_{i=1}^{n} \text{Var}(x_i) = (1/n) \sum_{i=1}^{n} \sum_{k=1}^{m} S_{ik}^2$$
$$= (1/n) \sum_{k=1}^{m} q_k^2 . \tag{2.2.23}$$

π^2 is the average, over predictors, of the predictor variance related to the principal predictors. It is a measure of the dependence of the predictors on the principal predictors. For brevity, π^2 will be called the average criterion-related predictor variance. This description should not imply that π^2 is an average multiple correlation of x predicted from y (roles of predictors and criteria reversed). π^2 is, however, the average multiple correlation of the x variables predicted from the principal predictors w_1. Note that π^2 is also the average of the q_k^2 (but the division is by n, not m, so that π^2 has a maximum value of 1.0).

Consider now two populations, each with the same average squared multiple correlation ρ^2. One population might have a small value of π^2 and the second a large value of π^2. In the first population the predictors depend very little on the principal predictors while in the second the dependence is greater. Nevertheless the prediction of the criteria is the same in each population. This paradoxical situation can be understood by first noting that we are considering for the moment prediction in the population, not in finite samples. The prediction of y is solely via the principal predictors w_1 and the random vector w_1 is an *exact* linear combination of the predictors x since S is square and non-singular:

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = S^{-1}x. \tag{2.2.24}$$

As additional verification that the average multiple correlation is independent of π^2, note that ρ^2 depends only on F in (2.2.17).

The parameter π^2 *will* have an effect on prediction in finite samples however. Consider prediction from the principal components of the predictors. When π^2 is large, the largest principal component will be largely in the space of the principal predictors and will contribute to prediction. However, when π^2 is small, the first principal component will be unrelated to the principal predictor space and will be a very poor predictor. The effectiveness of prediction from the principal components will thus depend on the size of π^2. Empirical confirmation of this phenomenon will be demonstrated in Section 3.4.

2.3 Computer Generation of the Model

Since any set of n predictors and m criteria may be written in the form (2.2.3), it is possible to generate an arbitrary population distribution by

specifying the matrices S, F, and E. The covariance matrices Σ_{xx}, Σ_{xy}, and Σ_{yy} may then be calculated by (2.2.4) to (2.2.6) where S is related to S_1 and S_2 by (2.2.19).

Rather than allowing the matrices S, F, and E to be completely arbitrary, a few basic parameters may be fixed arbitrarily and the matrices then generated essentially randomly subject to these given parameters. These arbitrary parameters are called "input parameters" since they are input to the computer program that generates the model.

The major input parameters are:

1. n = number of predictors.
2. m = number of criteria.
3. $(D_{kk}^2, k = 1, \cdots, m)$, the eigenvalues of $\Sigma_{\hat{y}\hat{y}}$.
4. $(q_k^2, k = 1, \cdots, m)$, the dependencies of the predictors on the principal predictors.

Note that once D_{kk}^2 and q_k^2 are specified for all k, the average squared multiple correlation ρ^2 and the average criterion-related predictor variance π^2 are fixed by (2.2.17) and (2.2.23). In particular, when m = 1 as in Chapter 3, $\rho^2 = D_{11}^2$ and $\pi^2 = (1/n)q_1^2$.

Two additional parameters, related to n and m, are:

1a. n_s = number of columns of S.
2a. m_d = number of duplicate criteria.

In the model described in Section 2.2, $S = (S_1 S_2)$ is an $(n \times n)$ square matrix. S_1 has m columns and S_2 has $(n - m)$ columns. In some of the experiments described in Sections 3.1 and 3.2, S has more than n columns, namely n_s columns, so that S_2 has $(n_s - m)$ columns and S_1 still has m columns.

All the experiments in Chapter 3 involve only one criterion (m = 1). However several *duplicate* criteria are allowed and the number of such duplicate criteria is denoted by m_d. Duplicate criteria are described further in Chapter 3.

Four minor parameters are needed to complete the input for the model:

5. v_x = variance of the generated $\mathrm{Var}(\hat{x}_i)$.
6. e_x = tolerance on this variance.
7. v_y = variance of the generated multiple correlations ρ_j^2.
8. e_y = tolerance on this variance.

The computer program generates matrices F and E so that the m squared multiple correlations ρ_j^2 calculated from them have mean exactly equal to the average of the D_{kk}^2 and *variance* equal to v_y within a maximum error of e_y. That is,

$$\rho^2 = (1/m) \sum_{j=1}^{m} \rho_j^2 = (1/m) \sum_{k=1}^{m} D_{kk}^2 \tag{2.3.1}$$

and

$$|\mathrm{Var}\,(\rho_j^2) - v_y| = \left|(1/m)\sum_{j=1}^{m}(\rho_j^2 - \rho^2)^2 - v_y\right| < e_y\,. \tag{2.3.2}$$

Matrix S, composed of S_1 and S_2, is generated so that the mean of the variances of $\hat{x}_j$ calculated from S_1 is exactly equal to $(1/n)\sum_{k=1}^{m} q_k^2$ and the variance of these variances is equal to v_x within a maximum error of e_x. That is,

$$\pi^2 = (1/n)\sum_{i=1}^{n}\mathrm{Var}\,(\hat{x}_i) = (1/n)\sum_{k=1}^{m} q_k^2 \tag{2.3.3}$$

and

$$|\mathrm{Var}\,[\mathrm{Var}\,(\hat{x}_i)] - v_x| < e_x\,. \tag{2.3.4}$$

The two occurrences of "Var" in the preceding formula refer to different types of variances. $\mathrm{Var}(\hat{x}_i)$ means the variance of the random variable $\hat{x}_i$ in the population. Let the constants $\mathrm{Var}(\hat{x}_i) = v_i$ temporarily. Then $\mathrm{Var}(v_i)$ is simply shorthand notation for

$$\mathrm{Var}\,(v_i) = (1/n)\sum_{i=1}^{n}(v_i - \pi^2)^2. \tag{2.3.5}$$

Note that π^2 is the mean of the v_i.

Generation of F

Since F is orthogonal by columns (equation (2.2.10)), it may be written as a product of an orthonormal matrix V and a diagonal matrix D consisting of the square roots of the eigenvalues D_{kk}^2:

$$F = VD. \tag{2.3.6}$$

The matrix D^2 is input so that generation of F reduces to the generation of an orthonormal V satisfying the two restrictions (2.3.1) and (2.3.2) on the squared multiple correlations ρ_j^2. ρ_j^2 may be expressed in terms of V and D by

$$\rho_j^2 = \mathrm{Var}\,(\hat{x}_j) = (FF')_{jj} = (VD^2V')_{jj} = \sum_{k=1}^{m} V_{jk}^2\, D_{kk}^2\,. \tag{2.3.7}$$

When ρ_j^2 is calculated in this way with V orthonormal, (2.3.1) is automatically satisfied. The restrictions on V are then

(a) V must be orthonormal,
(b) all ρ_j^2 calculated from (2.3.7) must be less than 1.0, and
(c) the variance of the ρ_j^2 must satisfy (2.3.2).

An algorithmic procedure to generate a V satisfying these three conditions, for arbitrary parameters m, (D_{kk}^2, k = 1, $\cdots$, m), v_y, and e_y, is outlined in Herzberg [1967].

Generation of E

The elements of the (m $\times$ m) matrix E are first generated randomly from N(0, 1) and then the rows of E are normalized so that, for all j,

$$(EE')_{jj} = \sum_{k=1}^{m} E_{jk}^2 = 1 - \rho_j^2 \tag{2.3.8}$$

where the ρ_j^2 are calculated from (2.3.7). This normalization ensures that the criteria y are normalized to unit variance. The methods used to generate normal random numbers as well as other random numbers discussed in this chapter follow Box and Muller [1958] and are described in Herzberg [1967].

Generation of S

Two different methods for generating S were developed for the experiments described in Chapters 3 and 4. The first is called the $e_x = 0$ method since the variance of $\mathrm{Var}(\hat{x}_i)$ is exactly equal to v_x. This method was used for the single criterion case (m = 1) in Chapter 3. The method does not generalize well to the m > 1 case and therefore a second method, called the $e_x \neq 0$ method was used for the several criterion calculations in Chapter 4. This latter method could have been employed in Chapter 3 except that e_x cannot be set to zero in this method.

When $m = 1$, π^2 can be calculated directly from the input parameters as

$$\pi^2 = (1/n)q_1^2 . \tag{2.3.9}$$

Then, in the $e_x = 0$ method, n numbers are generated randomly from $N(\pi^2, v_x)$ subject to the restriction that no number be more than 1.0 or less than 0.0. The numbers are rescaled after generation so that their mean is exactly π^2 and their variance is v_x. If one of the numbers is now more than 1.0 or less than 0.0, the number is discarded and a new attempt is made to satisfy the conditions. These n numbers are the variances of $\hat{x}_i$, denoted by $v_i = \mathrm{Var}(\hat{x}_i)$ in (2.3.5). When this step is completed, (2.3.3) and (2.3.4) are exactly satisfied with $e_x = 0$.

Now S_1 is a single column, s_1, and its elements are defined as the square roots of (v_i, i = 1, $\cdots$, n), with their signs chosen randomly. The elements of the last ($n_s - 1$) columns of S, namely S_2, are generated randomly from N(0, 1) and then rescaled, by rows, so that the row sum of squares of the whole S matrix is unity. This rescaling ensures that all x variables have unit variance. The generation of S by the $e_x = 0$ method is now complete.

The $e_x \neq 0$ method of generating S is very similar to the method of generating F. In analogy to (2.3.6) in which all matrices are (m $\times$ m), S_1 is

written as

$$S_1 = TQ \tag{2.3.10}$$

where S_1 and T are (n × m) and Q is diagonal (m × m) with diagonal elements = (q_k, k = 1, ⋯ , m). T (like V) is orthonormal by columns. The Var($\hat{x}_i$) may be written in terms of T and Q as

$$\text{Var}(\hat{x}_i) = (S_1 S_1')_{ii} = (TQ^2T')_{ii} = \sum_{k=1}^{m} T_{ik}^2 q_k^2 . \tag{2.3.11}$$

Since T is orthonormal it follows that the average of these variances is exactly $(1/n) \sum_{k=1}^{m} q_k^2 = \pi^2$ so that (2.3.3) is exactly satisfied. The remaining restrictions on T (as on V) are then

(a) T must be orthonormal by columns,
(b) all Var($\hat{x}_i$) calculated from (2.3.11) must be less than 1.0, and
(c) the variance of the Var($\hat{x}_i$) must satisfy (2.3.4).

The algorithmic procedure for generating V may be also used to generate a T satisfying the above three conditions for arbitrary n, m, (q_k^2, k = 1, ⋯ , m), v_x , and e_x .

After S_1 is generated, S_2(n × (n_s − m)) is generated in the same way as in the e_x = 0 method. First the elements of S_2 are generated as N(0, 1) random numbers. The rows of S_2 are rescaled so that each row sum of squares of S = (S_1S_2) is unity, resulting in unit variance x variables. This completes the $e_x \neq 0$ method for generating S.

2.4 Computer Generation of Data Samples

x and y are (n + m) random variables with a joint multivariate normal distribution. The mean vector is the null vector and the covariance matrix is Σ as given in (2.2.1). In order to draw samples from this distribution, it would be a straightforward procedure to use (2.2.3) which expresses x and y in terms of independent N(0, 1) variables w. For each simulated subject it would be necessary to generate (n + m) independent N(0, 1) numbers and to place these in (2.2.3) as the w values. The sample x and y vectors would then be found by matrix multiplication.

This procedure, while conceptually simple, has the disadvantage that that the computer time required increases linearly with N, the sample size. The method is thus costly to use for all but small sample sizes.

Another procedure was chosen instead. It is not based on generating sample vectors x and y at all but on generating a sample covariance matrix

$$C = \begin{bmatrix} C_{xx} & C_{yx} \\ C_{xy} & C_{yy} \end{bmatrix}. \tag{2.4.1}$$

The method is the Bartlett decomposition of the Wishart distribution [Bartlett, 1933; Kshirsagar, 1959; Wijsman, 1957]. The covariance matrix C has a Wishart distribution depending solely on the population covariance matrix Σ, the sample size N, and the number of variables which is $(n + m)$.

Let the population covariance matrix Σ be written as

$$\Sigma = \Omega\Omega'. \tag{2.4.2}$$

This may be done in a variety of ways. The Gauss-Doolittle method for computing a triangular Ω was used.

Let an $((n + m) \times (n + m))$ matrix A be defined as

$$A = (1/N)TT' \tag{2.4.3}$$

where T is a lower triangular $((n + m) \times (n + m))$ matrix whose lower triangular elements are independent random variables:

$$\begin{aligned} &T_{ij}(i > j) \quad \text{are} \quad N(0, 1) \\ &T_{ii} \quad \text{are} \quad \chi(N - i) \\ &T_{ij} = 0 \qquad (i < j). \end{aligned} \tag{2.4.4}$$

Then, if we compute

$$C = \Omega A\Omega' = (1/N)\Omega TT'\Omega', \tag{2.4.5}$$

C will have a Wishart distribution as desired. Equation (2.4.5) is the Bartlett decomposition of the Wishart matrix C. A is a sample covariance matrix from a population with identity covariance matrix. The letter A is used as a temporary symbol in this paragraph and is reserved for another use in Section 2.6.

2.5 Multiple Regression on the Predictors

The most widely used method for prediction is multiple regression. This least squares method ensures that, in the derivation sample, the correlation between the predicted score and the observed criterion score is a maximum. The maximum correlation is the multiple correlation.

Let the covariance matrix of the n predictors in the derivation sample (of size N) be $C_{xx}(n \times n)$ and let the column vector $c_{xy}(n \times 1)$ be the covariance between each predictor and the single criterion $(m = 1)$. Let the coefficients of the multiple regression combination of the predictors be $b_1(n \times 1)$, the subscript indicating that the first method of prediction, multiple regression on the predictors, is being used. The linear combination of the predictors is then b_1x.

The solution for b_1 is well known to be

$$b_1 = C_{xx}^{-1}c_{xy}. \tag{2.5.1}$$

The correlation between $\hat{y} = b_1'x$ and y is the multiple correlation. The square of this correlation is

$$r_1^2 = \frac{b_1'c_{xy}}{c_{yy}} \tag{2.5.2}$$

where c_{yy} is the sample variance of the criterion. The subscript on r^2 is used only in this chapter to distinguish the three methods of prediction. The subscript is dropped in later chapters.

2.6 Multiple Regression on the Principal Components

As an alternative to the original predictors one can use the largest principal components of the predictors in the regression function. The scores on the principal components are first estimated from the predictor scores.

The following calculations are made in the derivation sample. The characteristic roots and vectors of the covariance matrix of the predictors, $C_{xx}(n \times n)$, are calculated. Let the roots be written in descending order on the diagonal of a diagonal matrix U^2 and the vectors in corresponding order as the columns of an orthonormal matrix W. Then C_{xx} may be written as

$$C_{xx} = WU^2W' \tag{2.6.1}$$

where all matrices are $(n \times n)$. The characteristic vectors are the coefficients for relating the principal components, f, to the predictors, i.e.

$$f = W'x \tag{2.6.2}$$

where x is the $(n \times 1)$ column vector of one subject's scores on the predictors and f is the $(n \times 1)$ column vector of the principal component scores [Anderson, 1958, pp. 273–277].

We wish to use the t largest principal components in the regression function. From (2.6.2), the scores on these t principal components are estimated by

$$_t(f) = {}_t(W')x \tag{2.6.3}$$

where $_t(f)$ is $(t \times 1)$ and is the vector of the first t elements of f. $_t(W')$ is the matrix consisting of the first t rows of W'.

The prediction equation, using the t largest principal components, is

$$\hat{y}^{(t)} = d' \, {}_t(f) \tag{2.6.4}$$

where d is a temporary symbol representing the t-vector of weights for the principal components. The multiple regression solution for d is

$$d = C_{ff}^{-1}c_{fy} \tag{2.6.5}$$

in analogy to the solution (2.5.1). In (2.6.5), C_{ff} is $(t \times t)$ and c_{fy} is $(t \times 1)$. The covariance matrix of the t principal components is, from (2.6.3), (2.6.1),

and the orthonormality of W,

$$C_{ff} = {}_t(W')C_{xx}(W)_t = {}_t(U^2)_t \tag{2.6.6}$$

and the covariance of the t principal components and the criterion is

$$c_{fy} = {}_t(W')c_{xy} . \tag{2.6.7}$$

Therefore the weight vector is

$$d = {}_t(U^{-2})_t \, {}_t(W')c_{xy} . \tag{2.6.8}$$

Substituting (2.6.3) and (2.6.8) into (2.6.4), we find that

$$\hat{y}^{(t)} = c'_{xy}(W)_t \, {}_t(U^{-2})_t \, {}_t(W')x \tag{2.6.9}$$

is the equation for predicting y from x using regression on the t largest principal components of the predictors. This equation may be simplified slightly by writing

$$A = WU \tag{2.6.10}$$

so that (2.6.1) becomes

$$C_{xx} = AA'. \tag{2.6.11}$$

Then (2.6.9) becomes

$$\hat{y}^{(t)} = c'_{xy}(A')_t^{-1} \, {}_t(A^{-1})x. \tag{2.6.12}$$

Equation (2.6.12) is the formula for predicting y from x using regression on the t largest principal components. If we express the right hand side of this equation as $[b_2^{(t)}]'x$, then the weight vector is

$$b_2^{(t)} = (A')_t^{-1} \, {}_t(A^{-1})c_{xy} . \tag{2.6.13}$$

It can be shown [Herzberg, 1967] that the squared multiple correlation is

$$[r_2^{(t)}]^2 = \frac{[b_2^{(t)}]'c_{xy}}{c_{yy}} \tag{2.6.14}$$

which is identical in form to (2.5.2) but of course involves $b_2^{(t)}$ instead of b_1. It is important to note that $r_2^{(t)}$ is *not* the multiple correlation of y with the predictors x. The latter multiple correlation is r_1. The symbol $r_2^{(t)}$ represents the multiple correlation of y and the t largest principal components of the predictors and therefore $r_2^{(t)}$ must be less than r_1 unless $t = n$. The subscript 2 is dropped in later chapters when the context makes clear which method of prediction is used.

Multiple regression on the principal components may be called a reduced rank method since the use of the t largest principal components instead of the original predictors is equivalent to approximating the matrix C_{xx} by the matrix

$$\tilde{C}_{xx} = (A)_{t\ t}(A'). \tag{2.6.15}$$

The matrix $\tilde{C}_{xx}$ is of reduced rank $t < n$.

2.7 Multiple Regression on the Principal Predictors

Another method of calculating independent scores in the derivation sample is the method of principal predictors. Scores on the largest principal predictors are used in the regression equation. The method is only applicable if there are several criteria ($m > 1$).

Let the scores of a subject on the m criteria be $y_1, \cdots, y_m$, which may be placed in a column vector y ($m \times 1$). Each criterion, y_j, has a part, $\hat{y}_j$, linearly predictable from the n predictors, x, in the derivation sample:

$$\hat{y}_j = c'_{xy\,j} C_{xx}^{-1} x \tag{2.7.1}$$

where c_{xy_j}($n \times 1$) is the covariance of y_j with the n predictors x and C_{xx} is the covariance of the predictors. The m predictable parts of the criteria may be written as a column vector $\hat{y}$($m \times 1$). Then (2.7.1) may be rewritten as

$$\hat{y} = C'_{xy} C_{xx}^{-1} x = C_{yx} C_{xx}^{-1} x \tag{2.7.2}$$

where C_{yx}($m \times n$) is the covariance matrix of the m criteria with the n predictors. The covariance of the predictable parts is the ($m \times m$) matrix $C_{\hat{y}\hat{y}}$:

$$C_{\hat{y}\hat{y}} = C_{yx} C_{xx}^{-1} C'_{yx}. \tag{2.7.3}$$

Let us diagonalize $C_{\hat{y}\hat{y}}$ in analogy to the way that C_{xx} was written in (2.6.1) and (2.6.11):

$$C_{\hat{y}\hat{y}} = VD^2V' = GG'. \tag{2.7.4}$$

The matrices V and D^2 are sample estimates of the corresponding population matrices denoted by the same symbols in Sections 2.2 and 2.3. The eigenvalues D^2_{kk} are written in decreasing order in the diagonal of D^2 and the eigenvectors are written in the corresponding order as columns of V. The rows of G are the coefficients for relating the predictable parts of the criteria and the principal predictors, i.e.,

$$\hat{y} = Gw. \tag{2.7.5}$$

Thus the equation for estimating the scores on the m principal predictors (column vector w ($m \times 1$)) from the predictable parts is

$$w = G^{-1}\hat{y} \tag{2.7.6}$$

which, combined with (2.7.2), gives

$$w = G^{-1} C_{yx} C_{xx}^{-1} x \tag{2.7.7}$$

as the equation for estimating w from x.

The equation for estimating the t largest principal predictors ${}_t(w)$ is

$$(2.7.8) \qquad {}_t(w) = {}_t(G^{-1})C_{yx}C_{xx}^{-1}x$$

where ${}_t(G^{-1})$ is the first t rows of G^{-1}.

Note that w is a vector of numbers which can be calculated for each subject. The use here of w for the estimated principal predictor score vector should not be confused with the use of w, in Section 2.2, for the principal predictors, a random vector in the population.

Since the principal predictors are uncorrelated in the derivation sample, the multiple regression weights for predicting each y_j from the principal predictors are simply the rows of the pattern matrix G as shown in (2.7.5). Each row of G is the weight vector for one criterion variable. The coefficients G are still the correct weights when only some of the principal predictors are included in the regression. If the t largest principal predictors are included, the predicted parts of the criteria are

$$(2.7.9) \qquad \hat{y}_3^{(t)} = (G)_t \, {}_t(w).$$

In order to express this equation in terms of the original predictor scores as

$$(2.7.10) \qquad \hat{y}_3^{(t)} = B_3'x,$$

(2.7.9) and (2.7.8) may be combined, yielding

$$(2.7.11) \qquad B_3 = C_{xx}^{-1}C_{xy}(G')_t^{-1} \, {}_t(G').$$

It can be shown [Herzberg, 1967] that the squared multiple correlation of the jth criterion variable y_j is

$$(2.7.12) \qquad [r_{3j}^{(t)}]^2 = \frac{(B_3'C_{xy})_{jj}}{c_{y_jy_j}}$$

which is identical in form to (2.5.2) and (2.6.14) except that there is one such equation for each criterion variable y_j. As was the case with regression on the principal components, the multiple correlation $r_{3j}^{(t)}$ is *not* the multiple correlation of y_j with x but is the multiple correlation of y_j and the t largest principal predictors. The correlation $r_{3j}^{(t)}$ is always less than or equal to r_1. The subscript 3 is dropped in later chapters.

Multiple regression on the principal predictors, as on the principal components, is a reduced rank method. The use of the t largest principal predictors instead of all m principal predictors is equivalent to approximating the matrix $C_{\hat{y}\hat{y}}$ by the matrix

$$(2.7.13) \qquad \tilde{C}_{\hat{y}\hat{y}} = (G)_t \, {}_t(G').$$

The matrix $\tilde{C}_{\hat{y}\hat{y}}$ is of reduced rank $t < n$.

It was pointed out after (2.7.4) that the eigenvalues of $C_{\phi\phi}$ are estimates of the population parameters (D_{kk}^2, k = 1, $\cdots$, m). In order to estimate π^2 and (q_k^2, k = 1, $\cdots$, m), it is natural to require that the covariance matrices be changed to correlation matrices. The correlations of the predictors x and the principal predictors w are given by the (n $\times$ m) sample matrix $S_1 = C_{xw}$. S_1 may be written as

$$S_1 = C_{xw} = C_{xx}C_{xx}^{-1}C_{xy}(G')^{-1} = C_{xy}(G')^{-1} \tag{2.7.14}$$

by equation (2.7.7).

The sample quantity q_k^2 is the total dependence of the predictors on the kth principal predictor and is therefore, since the predictors have unit variance by the use of correlation matrices, given by

$$\text{sample } q_k^2 = \sum_{i=1}^{n} S_{ik}^2 . \tag{2.7.15}$$

Similarly the sample estimate of π^2, the average criterion-related predictor variance, is the sum of the squares of all the elements of S_1 divided by n and is

$$\begin{aligned} \text{sample } \pi^2 &= (1/n) \sum_{k=1}^{m} \sum_{i=1}^{n} S_{ik}^2 \\ &= (1/n) \sum_{k=1}^{m} \text{sample } q_k^2 . \end{aligned} \tag{2.7.16}$$

2.8 Cross-Validities

The calculation of correlations in the validation sample is identical for all three weight computational methods. Given the weight vector b from the derivation sample, the square of the correlation between $b'x$ and a criterion variable y in the validation sample is

$$r_c^2 = \frac{(b'c_{xy})^2}{(b'C_{xx}b)c_{yy}} \tag{2.8.1}$$

where c_{xy}, C_{xx} and c_{yy} are covariances computed in the validation sample. The quantity r_c is called the *sample cross-validity* and may be negative or positive. The sign of r_c is equal to the sign of $b'c_{xy}$. The sign of r_c is also affixed to r_c^2 when averages of several r_c^2 are taken.

The predictor weight vector b, calculated in the derivation sample, may also be applied to the population itself. The square of the correlation between $b'x$ and the criterion variable y in the population is

$$\rho_c^2 = \frac{(b'\sigma_{xy})^2}{(b'\Sigma_{xx}b)\sigma_{yy}}. \tag{2.8.2}$$

This formula is identical to (2.8.1) except for the use of population covariances instead of sample covariances. ρ_c is called the *population cross-validity*. A sign is affixed to ρ_c^2 in the same way as to r_c^2.

The three statistics, r^2 (in its several forms), r_c^2, and ρ_c^2 are called the *correlation statistics*. These statistics are the principal quantities computed in the experiments described in Chapters 3 and 4.

CHAPTER 3

SIMULATION RESULTS WITH ONE CRITERION VARIABLE

A number of experiments was performed using a single criterion variable. The two methods of prediction used were multiple regression on the predictors and on the principal components of the predictors. The model and sample generation methods described in Sections 2.2 and 2.3 are applicable to the single criterion case ($m = 1$). However, in order to generate many samples without excessive use of computer time, a special procedure for the $m = 1$ case was developed. This procedure, described in detail in Herzberg [1967], allows any number of criterion variables to be generated, each with the same population multiple correlation and each with the same relation to the predictors. The criterion variables are thus all duplicates of the single criterion of interest. The number of such duplicates is denoted by m_d.

As an example, suppose that there are five predictors and ten duplicate criteria. Then, in a sample, one can compute ten multiple correlations, one for each criterion. These ten correlations are all based on one sample (size N) of five predictor scores but on ten different samples of single criterion scores.

Each calculation described in this chapter is based on one or more populations (Σ) generated for each combination of the input parameters. For each Σ that was generated, sample covariance matrices were generated in pairs (C_1, C_2), representing the two samples needed for double cross-validation. Σ, C_1, and C_2 are the covariance matrices of $(n + m_d)$ variables—n predictors and m_d duplicate criteria. Except in Section 3.4, at least two such pairs of sample covariance matrices were generated. This allowed variation in the predictor sample covariance matrices.

The results using the simulation program described in this chapter are:

(a) When $\rho = 0$, the sample multiple correlation and cross-validity follow the known theoretical law. (Section 3.1)
(b) When $\rho \neq 0$, variation in Σ_{xx} produced by change in the number of columns of S does not affect the correlation statistics r^2, r_c^2, and ρ_c^2. (Section 3.2)
(c) The correlation statistics depend on n, N, and ρ^2 in theoretically understandable ways. Tables are presented which may be used to interpret sample multiple correlations and cross-validities. (Section 3.3)
(d) The optimum number of principal components to include in the regression function depends on the parameters n, N, ρ^2, and π^2. (Section 3.4)

3.1 Distribution of the Correlation Statistics When $\rho = 0$

It is useful to study populations in which the multiple correlation (ρ) is zero even though such populations are of little practical significance. Firstly, the distributions of the correlation statistics when $\rho = 0$ provide a baseline against which to compare the distributions obtained for non-zero ρ. Secondly, some properties of the distributions for $\rho = 0$ are known theoretically and a comparison of the distributions obtained from the computer model for $\rho = 0$ with the theoretical predictions provides a check of the model and the computer calculations.

Fisher [1928] showed that, if r is the sample multiple correlation,

$$\frac{r^2}{1 - r^2} \frac{N - n - 1}{n} \tag{3.1.1}$$

is distributed as $F(n, N - n - 1)$ when $\rho = 0$. This distribution has the important property that it is independent of the covariance matrix of the predictors, Σ_{xx}. The distribution of the sample cross-validity, r_c, is the distribution of the sample correlation of two variables whose population correlation is zero. Therefore, from (3.1.1),

$$r_c\left(\frac{N - 2}{1 - r_c^2}\right)^{1/2} \tag{3.1.2}$$

is distributed as $t(N - 2)$ when $\rho = 0$. This distribution is also independent of Σ_{xx}. Finally, the population cross-validity, ρ_c, is exactly zero since $\sigma_{xy} = 0$.

An effective simulation of multiple regression should be able to reproduce these properties. The distributions of r^2 and r_c^2 were studied for two different sample sizes, N, and for predictor covariance matrices, Σ_{xx}, varying in two ways, namely in the values of the parameters π^2 and n_s. π^2 is the average of the variances of the part of each predictor dependent on the principal predictor; n_s is the number of columns of the S matrix. Two values of n_s were used—$n_s = 10$ implies a square S matrix since $n = 10$ and $n_s = 20$ implies a non-square S matrix.

The input parameters which were constant for all the calculations in this section are shown in Table 1. Table 2 shows the variable model parameters and sample sizes and also the total number of populations and samples calculated for each model. According to Table 2 four different Σs were generated for each model, and for each Σ generated, three double cross-validations were performed. Since ten duplicate criteria were used throughout ($m_d = 10$), there were altogether $10 \times 4 \times 3 \times 2 = 240$ sample multiple correlations and cross-validities computed for each model. The final factor of two in the preceding expression represents the two samples (C_1, C_2) which were generated for each double cross-validation.

In order to test whether r^2 satisfies the Fisher distribution law (3.1.1), it is necessary to have a tabulation of $F(10, 39)$ for Models 1 and 2 and

TABLE 1

Constant Parameters for Section 3.1

$n = 10$
$m_d = 10$ $(m = 1)$
$\rho^2 = 0.0$
$v_x = 0.01$
$e_x = 0.0$
v_y, e_y inapplicable since $m = 1$

F(10, 130) for Models 3 and 4. These distributions were obtained directly, or by interpolation, from Owen [1962]. From (3.1.1), r^2 is distributed as

$$\frac{nF(n, N - n - 1)}{(N - n - 1) + nF(n, N - n - 1)}. \tag{3.1.3}$$

The percentage points used (those available in Owen) and the corresponding percentile points of the F and r^2 distributions are shown in Tables 3 and 4.

The four Σs and associated samples for each model were divided equally into two sets, chosen in the order that they were computed. The cumulative frequency distribution of the 120 sample squared multiple correlations, r^2, are presented in Tables 3 and 4. Each set of 120 squared multiple correla-

TABLE 2

Variable Parameters for Section 3.1

	Model 1	Model 2	Model 3	Model 4
n_s	10	20	10	20
π^2	.2	.2	.5	.5
number of Σs	4	4	4	4
N	50	50	131	131
number of C_1, C_2 pairs per Σ	3	3	3	3

TABLE 3

Distribution of r^2 for Models 1 and 2

Probability	F(10, 39)	r^2	Cumulative Frequencies				
				Model 1		Model 2	
			Expected	Set 1	Set 2	Set 1	Set 2
1.000		1.000	120	120	120	120	120
.975	2.401	.381	117	117	116	117	117
.95	2.086	.348	114	115	114	115	114
.90	1.769	.312	108	103	107	108	112
.75	1.329	.254	90	85	80	98	91
.50	.951	.196	60	50	52	60	59
.25	.664	.145	30	25	23	30	26
.10	.469	.107	12	16	12	11	9
.05	.375	.0878	6	5	6	5	6
.025	.307	.0729	3	1	3	4	4
.000	.000	.0000	0	0	0	0	0
MD = maximum absolute difference (observed - expected)				10	10	8	4
Kolmogorov-Smirnov D = MD/120				.083	.083	.067	.033

TABLE 4

Distribution of r^2 for Models 3 and 4

Probability	F(10,120)	r^2	Cumulative Frequencies				
				Model 3		Model 4	
			Expected	Set 1	Set 2	Set 1	Set 2
1.000		1.0000	120	120	120	120	120
.975	2.157	.1524	117	120	115	119	117
.95	1.911	.1373	114	119	111	113	113
.90	1.652	.1210	108	114	107	106	109
.75	1.279	.0963	90	95	96	88	95
.50	.939	.0726	60	65	69	60	69
.25	.670	.0529	30	27	39	33	32
.10	.480	.0385	12	7	18	13	13
.05	.388	.0313	6	6	9	7	5
.025	.318	.0259	3	4	2	5	3
.000	.000	.0000	0	0	0	0	0
MD = maximum absolute difference (observed - expected)				6	9	3	9
Kolmogorov-Smirnov D = MD/120				.050	.075	.025	.075

tions originated in six sample covariance matrices for each of two Σs. There were ten duplicate criteria in each sample.

The observed frequency distributions were compared with the expected distributions by the Kolmogorov-Smirnov one sample test [Siegel, 1956]. MD is the maximum absolute difference between observed and expected cumulative frequencies and D = MD/120 is the Kolmogorov–Smirnov statistic. Both values are presented in the tables. The critical D for the one sample, two tailed test is 0.12 ($\alpha = 0.05$, 120 observations). None of the Ds in Tables 3 and 4 exceeds this value. There is therefore good evidence that the multiple correlations generated by the simulation program satisfy the Fisher law.

The observed distributions of r_c were compared with the expected distributions. From (3.1.2), r_c is distributed as

$$\frac{t(N-2)}{(N-2+[t(N-2)]^2)^{1/2}}. \tag{3.1.4}$$

The percentile points of r_c are shown in Tables 5 and 6. In these tables the distributed variable is, for simplicity, r_c, not r_c^2, in order to emphasize that r_c is distributed about zero. If, consistent with all the other frequency tables, r_c^2 (with the sign of r_c) were chosen, the distribution would, of course, be

TABLE 5

Distribution of r_c for Models 1 and 2

Probability	t(48)	r_c	Cumulative Frequencies				
				Model 1		Model 2	
			Expected	Set 1	Set 2	Set 1	Set 2
1.000			120	120	120	120	120
.975	2.012	.2789	117	118	119	116	114
.95	1.678	.2354	114	109	116	114	111
.90	1.300	.1855	108	96	112	102	103
.75	.680	.0977	90	81	86	85	91
.50	.000	.0000	60	54	55	59	61
.25	-.680	-.0977	30	28	23	32	30
.10	-1.300	-.1855	12	15	3	11	16
.05	-1.678	-.2354	6	11	1	7	8
.025	-2.012	-.2789	3	7	1	4	5
.000			0	0	0	0	0
MD = maximum absolute difference (observed - expected)				12	9	6	5
Kolmogorov-Smirnov D = MD/120				.100	.075	.050	.042

TABLE 6

Distribution of r_c for Models 3 and 4

Probability	t(129)	r_c	Cumulative Frequencies				
				Model 3		Model 4	
			Expected	Set 1	Set 2	Set 1	Set 2
1.000			120	120	120	120	120
.975	1.979	.1716	117	119	114	118	116
.95	1.657	.1444	114	116	110	114	114
.90	1.288	.1127	108	114	101	109	112
.75	.677	.0595	90	103	82	99	97
.50	.000	.0000	60	73	49	77	76
.25	-.677	-.0595	30	37	22	37	35
.10	-1.288	-.1127	12	14	11	13	16
.05	-1.657	-.1444	6	5	4	6	10
.025	-1.979	-.1716	3	2	2	4	5
.000			0	0	0	0	0
MD = maximum absolute difference (observed - expected)				13	11	17	16
Kolmogorov-Smirnov D = MD/120				.108	.082	.142	.133

identical. The expected and observed frequency distributions and the Kolmogorov–Smirnov statistics are presented in these tables. The critical D (=0.12, as before) is exceeded in two cases. This result might be interpreted as failure of the simulation to produce truly independent derivation and validation samples. However, in 15 further models described in Section 3.3, all with $\rho^2 = 0.0$ and varying n and N, none of the 15 Kolmogorov–Smirnov Ds was significant at the .05 level. (In these models only 40 cross-validities were obtained for each Σ). Therefore it may be concluded that the two significant Ds in the present section are chance results.

The calculations in this section have shown that when the population multiple correlation is zero, the sample statistics obey the theoretically known distributions. The function (3.1.1) of the sample multiple correlation has an F distribution and the function (3.1.2) of the sample cross-validity has a t distribution.

3.2 Dependence of the Correlation Statistics on Σ_{xx} When $\rho \neq 0$

It was shown in Section 2.2 that the average squared population multiple correlation, ρ^2, depends only on the D^2_{kk} (equation (2.2.17)) and not on matrix S. In particular, when there is only one criterion (m = 1), the squared population multiple correlation of the criterion with the predictors is

(3.2.1) $$\rho^2 = D^2_{11} .$$

Since D_{11}^2 is an input parameter, it is therefore straightforward to specify an arbitrary ρ^2 for a desired population.

Unfortunately this simple specification of ρ^2 is only possible when the matrix S is square. When S is non-square, say with n_s columns, then ρ^2 is given by

$$\rho^2 = D_{11}^2 s_1'(SS')^{-1}s_1 \tag{3.2.2}$$

which depends on the matrix S. The vector s_1 is the first column of S (equation 2.2.19). Since S is generated to some extent randomly by the population generation procedure, it is impossible to specify by input parameters what the population multiple correlation will be. This is a severe limitation on the model if S is not square.

Because of the ease in specifying ρ^2, the models in the remaining sections of this study all employ square S matrices. In order to show, at least to a certain extent, that this does not effect the generality of the conclusions, some experiments are described in this section which compare the correlation statistics of square S and non-square S models. A further comparison is made of two models which differ only in the parameter π^2.

The input parameters which were constant for all the calculations in this section are shown in Table 7. Table 8 shows the variable parameters and the total number of populations and samples calculated for each of the ten models. Models 5 and 6 are square S models differing only in π^2. Two populations were generated for each model and three sample pairs for each population. This results in $10 \times 2 \times 3 \times 2 = 120$ sample correlations for each model.

Each of the four remaining model pairs differs only in n_s ; one model of each pair has square S and the other model has non-square S. The squared population multiple correlation, ρ^2, is identical for the model in each pair. The identity holds to six or seven decimal places even though only three are

TABLE 7

Constant Parameters for Section 3.2

$n = 5$
$m_d = 10$ $(m = 1)$
$v_x = 0.01$
$e_x = 0.00$
v_y, e_y inapplicable since $m = 1$
$N = 40$

TABLE 8

Variable Parameters for Section 3.2

	Model									
	5	6	7	8	9	10	11	12	13	14
n_s	5	5	5	10	5	10	5	10	5	10
ρ^2	.5	.5	.234	.234	.311	.311	.438	.438	.495	.495
π^2	.2	.5	.2	.2	.2	.2	.5	.5	.5	.5
number of Σs	2	2	1	1	1	1	1	1	1	1
number of C_1, C_2 pairs per Σ	3	3	6	6	6	6	6	6	6	6

shown in Table 8. The identity was produced in the following way. As explained above, ρ^2 is not an input parameter. The input D_{11}^2 is equal to ρ^2 only for square S models. The non-square S Models 8, 10, 12, and 14 were generated using $D_{11}^2 = 0.5$. The population squared multiple correlation, ρ^2, was calculated for each model by (3.2.2). These are the values in Table 8. These computed values were used as the input D_{11}^2 for the square S Models 7, 9, 11, and 13. Since only one Σ could be generated for each set of parameters, six C_1 , C_2 pairs were generated instead of three. This means that $10 \times 1 \times 6 \times 2 = 120$ sample correlations were generated, the same number as for Models 5 and 6.

The frequency distributions of the 120 correlations for each model are presented in Tables 9 (r^2), 10 (r_c^2), and 11 (ρ_c^2). Recall, from Section 2.8, that a negative r_c^2 results from a negative r_c . The maximum absolute difference, MD, and the Kolmogorov–Smirnov statistic D are also presented for each pair of models. The critical D for the two sample test, two tailed, is 0.18 ($\alpha = 0.05$, 120 observations). None of the sample values exceeds this value although two approach it. It can be safely stated that for these examples the variation in Σ_{xx} has not produced differences in the observed distributions of the correlation statistics.

In the simulation model, the population multiple correlation is independent of π^2, the average predictor variance related to the criteria. The comparison of Models 5 and 6 confirmed that the sample statistics are independent of π^2. Even though the population multiple correlation does depend on the matrix S if it is not square, it was shown that the correlation statistics are not affected by the matrix S for the models considered in this section. All further calculations in this study employ square S models only.

3.3 Dependence of the Correlation Statistics on n, N, *and* ρ^2

The cross-validation technique was developed as a way to correct a sample multiple correlation [Mosier, 1951]. The purpose of the calculations

TABLE 9

Cumulative Frequency Distribution of r^2 for Models 5 to 14

r^2	Model 5	6	7	8	9	10	11	12	13	14
.85								120	120	120
.80	120	120					120	119	119	119
.75	119	118					119	119	117	115
.70	108	111			120	120	115	119	113	109
.65	101	93	120		119	119	102	114	94	98
.60	76	68	119	120	116	115	90	97	68	71
.55	44	48	118	117	111	111	73	73	49	51
.50	31	29	115	113	99	96	48	50	27	29
.45	18	17	104	105	69	77	30	30	13	19
.40	8	4	89	92	53	60	13	20	5	10
.35	3	1	74	85	28	41	4	12	4	3
.30	1	0	45	66	15	28	2	7	1	1
.25	0		26	46	10	21	1	4	1	0
.20			11	21	4	10	0	1	0	
.15			9	12	1	6		0		
.10			2	3	0	1				
.05			0	1		0				
.00				0						
MD	8		21		13		12		6	
D	.067		.175		.109		.100		.050	

described in this section was to investigate empirically the relationship of the sample multiple correlation and the cross-validities to the population multiple correlation.

The parameters which were constant for all the calculations in this section are shown in Table 12. Table 13 indicates the values of the three parameters which were varied. All possible combinations of squared multiple correlations, ρ^2, number of predictors, n, and sample size, N, were used except for those combinations with (n, N) = (15, 16). A different model was generated for each combination of (ρ^2, n, N) so that the models for combinations differing only in N are different models.

In all cases 40 sample correlations were obtained. The means of the 40 correlations of each type (r^2, r_c^2, and ρ_c^2) are shown in Tables 14 to 17. The expected values $E(r^2)$ as calculated from (1.2.10) are also shown in these tables. The standard errors of the correlation means range from 0.01 to 0.02 for r^2 and r_c^2 and somewhat less for ρ_c^2. However, this standard error does

TABLE 10

Cumulative Frequency Distribution of r_c^2 for Models 5 to 14

	Model									
r_c^2	5	6	7	8	9	10	11	12	13	14
.75		120							120	120
.70	120	119							119	117
.65	119	115			120		120	120	115	115
.60	112	105			119		112	114	110	106
.55	102	83		120	119	120	99	108	91	93
.50	72	68	120	119	116	118	85	94	66	71
.45	55	56	118	118	113	110	74	82	51	47
.40	38	29	112	113	104	100	59	61	26	32
.35	27	18	106	111	88	86	43	42	13	24
.30	14	10	99	102	64	68	25	26	8	15
.25	8	4	88	98	44	53	13	17	4	8
.20	2	2	66	81	22	35	7	7	2	3
.15	0	0	54	52	12	25	5	3	0	0
.10			31	36	7	18	3	0		
.05			15	16	3	6	0			
.00			2	1	0	2				
-.05			0	1		0				
-.10				0						
MD	19		15		13		9		11	
D	.158		.125		.109		.075		.092	

not take into account variation which would be produced by another population generated from the same parameters. Table 13 indicates that only one Σ was generated for each parameter combination. Nevertheless the tables give a useful picture of the dependence of the three correlation statistics on ρ^2, N, and n.

The following observations may be made from the tables:

(a) The squared sample multiple correlation, r^2, is an overestimate of ρ^2. The expected values of r^2, $E(r^2)$, from (1.2.10) match the observed r^2 values very well. The first two terms of the expansion (1.2.11) may be rearranged to show that the bias in $E(r^2)$ is a simple function of n, N, and ρ^2:

$$E(r^2) - \rho^2 = \frac{n}{N-1}(1 - \rho^2). \tag{3.3.1}$$

The match of $E(r^2)$ and r^2 shows that formulas (1.2.13) and (1.2.7),

which are essentially backward solutions of (1.2.10), provide reasonable estimates of the squared population multiple correlation.

(b) The squared sample cross-validity, r_c^2, is generally an underestimate of ρ^2 and this bias (except for $\rho^2 = 0.0$) tends to be approximately the same for all values of ρ^2 for fixed (n, N). As with r^2, the bias of the cross-validity decreases with N and increases with n for fixed ρ^2.

(c) The squared population cross-validity, ρ_c^2, is the squared cross-validity using the derived weights on a validation sample of infinite size. The tables show that ρ_c^2 has similar values to r_c^2 and the same comments apply to ρ_c^2 as were applied to r_c^2 above. Looked at another way, r_c^2 is an unbiased estimate of ρ_c^2.

TABLE 11

Cumulative Frequency Distribution of ρ_c^2 for Models 5 to 14

	Model									
ρ_c^2	5	6	7	8	9	10	11	12	13	14
.500	120	120							120	120
.475	82	93							94	99
.450	53	56					120	120	59	61
.425	31	33					113	107	29	26
.400	18	16					78	71	18	6
.375	13	7					47	37	9	3
.350	8	3					28	21	4	1
.325	6	2			120	120	14	11	2	0
.300	3	0			108	114	9	6	2	
.275	1				82	81	4	3	0	
.250	1		120	120	57	52	1	1		
.225	1		118	113	24	32	1	0		
.200	1		88	91	14	14	1			
.175	1		65	65	5	10	0			
.150	1		37	45	2	5				
.125	0		18	32	2	3				
.100			9	15	1	3				
.075			4	10	0	1				
.050			3	4		1				
.025			2	4		1				
.000			0	1		0				
-.025				0						
MD	11		14		8		10		12	
D	.092		.117		.067		.083		.100	

TABLE 12

Constant Parameters for Section 3.3

$m = 1$

$n_s = n$ (square S matrix)

$\pi^2 = 0.5$

$v_x = 0.01$

$e_x = 0.0$

v_y, e_y inapplicable since $m = 1$

The tables confirm the known properties of multiple regression and cross-validation as summarized in (1.2.14). The sample multiple correlation is an overestimate of the population value since the sample weights are chosen to optimize the correlation in the derivation sample. These weights are not the optimum weights in either the population or another sample and the consequence is that both ρ_c^2 and r_c^2 are biased low. With repeated samplings the values of r_c^2 cluster around ρ_c^2 since some weights are better for the validation sample than the population ($r_c^2 > \rho_c^2$) and other weights are worse ($r_c^2 < \rho_c^2$).

These tables can, perhaps, be useful in estimating the population ρ^2 from sample r^2 and r_c^2 values obtained from real data. If the values of (n, N)

TABLE 13

Variable Parameters for Section 3.3

$n = 2, 5, 10, 15$

$\rho^2 = 0.0, 0.1, 0.25, 0.5, 0.75$

$N = 16, 26, 50, 131$

	models with (n, N) =		
	(15, 16)	(10, 16)	all others
m_d		5	10
number of Σs	not done	1	1
number of C_1, C_2 pairs per Σ		4	2

correspond to one of the tables and if r^2 and r_c^2 match the values in the tables for a value of ρ^2, then this value of ρ^2 is the estimate of the population multiple correlation.

3.4 Prediction from the Principal Components

Burket [1964] showed that cross-validities can be increased by using only a few principal components of the predictors in the prediction function. The

TABLE 14

Correlation Statistics for Sample Size N = 16

ρ^2	.00	.10	.25	.50	.75
			n = 2		
$E(r^2)$	.133	.211	.331	.541	.763
r^2	.160	.189	.329	.593	.805
r_c^2	.029	.121	.217	.534	.763
ρ_c^2	.000	.056	.191	.468	.731
			n = 5		
$E(r^2)$	.333	.393	.485	.647	.818
r^2	.307	.409	.450	.669	.866
r_c^2	.014	.086	.104	.372	.691
ρ_c^2	.000	.027	.107	.351	.648
			n = 10		
$E(r^2)$	.667	.696	.743	.823	.909
r^2	.680	.666	.744	.821	.908
r_c^2	-.035	.003	.047	.300	.523
ρ_c^2	.000	.009	.041	.217	.489

TABLE 15

Correlation Statistics for Sample Size N = 26

ρ^2	.00	.10	.25	.50	.75
			n = 2		
$E(r^2)$	.080	.166	.297	.523	.757
r^2	.073	.183	.372	.542	.805
r_c^2	.016	.125	.317	.491	.788
ρ_c^2	.000	.076	.226	.480	.738
			n = 5		
$E(r^2)$	.200	.275	.389	.585	.789
r^2	.188	.277	.400	.565	.733
r_c^2	-.013	.059	.204	.384	.642
ρ_c^2	.000	.046	.161	.413	.710
			n = 10		
$E(r^2)$	.400	.456	.542	.689	.841
r^2	.409	.455	.567	.685	.855
r_c^2	-.010	.012	.145	.335	.642
ρ_c^2	.000	.016	.102	.311	.629
			n = 15		
$E(r^2)$	.600	.637	.694	.792	.894
r^2	.604	.648	.669	.804	.888
r_c^2	.009	.030	.050	.252	.425
ρ_c^2	.000	.012	.050	.199	.476

TABLE 16

Correlation Statistics for Sample Size N = 50

ρ^2	.00	.10	.25	.50	.75
		n = 2			
$E(r^2)$	.041	.133	.274	.511	.753
r^2	.037	.139	.276	.524	.738
r_c^2	.002	.107	.239	.496	.733
ρ_c^2	.000	.079	.233	.485	.746
		n = 5			
$E(r^2)$	.102	.189	.320	.542	.769
r^2	.092	.182	.350	.547	.765
r_c^2	.007	.079	.218	.468	.720
ρ_c^2	.000	.057	.198	.453	.725
		n = 10			
$E(r^2)$	.204	.281	.397	.594	.795
r^2	.202	.263	.399	.634	.790
r_c^2	.005	.063	.137	.476	.694
ρ_c^2	.000	.042	.136	.418	.693
		n = 15			
$E(r^2)$	.306	.373	.474	.646	.821
r^2	.327	.353	.483	.659	.806
r_c^2	-.003	.033	.160	.350	.666
ρ_c^2	.000	.027	.116	.350	.671

TABLE 17

Correlation Statistics for Sample Size N = 131

ρ^2	.00	.10	.25	.50	.75
			n = 2		
$E(r^2)$	.015	.113	.259	.504	.751
r^2	.014	.118	.241	.526	.764
r_c^2	.001	.101	.235	.515	.761
ρ_c^2	.000	.091	.246	.496	.748
			n = 5		
$E(r^2)$	.038	.133	.276	.516	.757
r^2	.040	.127	.299	.500	.769
r_c^2	-.001	.075	.245	.465	.749
ρ_c^2	.000	.076	.227	.483	.742
			n = 10		
$E(r^2)$	.077	.168	.305	.534	.767
r^2	.085	.178	.295	.537	.767
r_c^2	.002	.078	.209	.473	.730
ρ_c^2	.000	.061	.205	.464	.730
			n = 15		
$E(r^2)$	.115	.203	.334	.554	.776
r^2	.116	.202	.325	.573	.799
r_c^2	-.000	.037	.174	.463	.753
ρ_c^2	.000	.048	.181	.449	.722

TABLE 18

Constant Parameters for Section 3.4

$m_d = 10$ $(m = 1)$
$n_s = n$ (square S matrix)
$v_x = 0.01$
$e_x = 0.0$
v_y, e_y inapplicable since $m = 1$
number of Σs per model = 1
number of C_1, C_2 pairs per $\Sigma = 1$

formulas for such prediction were presented in Section 2.6. The calculations described in the present section demonstrate the improvement of prediction by using the largest principal components in the simulation data and show how variation in some parameters can change the effect.

Sixty models were generated varying in four parameters: n, the number of predictors; ρ^2, the squared multiple correlation; π^2, the average criterion-related predictor variance; and N, the sample size. The constant parameters are listed in Table 18. All combinations of the variable parameters listed in Table 19 were used. A new model was generated for each (n, ρ^2, π^2, N) combination so that here, as in Section 3.3, combinations differing only in N are different models.

For each simulation model two sample covariance matrices, C_1 and C_2, were generated, both with the same sample size N. The covariance matrix of the first sample predictors, $(C_1)_{xx}$, was diagonalized and the weights for predicting each of the ten duplicate criteria from the largest principal components were calculated. These weights were validated on C_2, and the cross-validities $r_c^{(1)}$ were calculated for each criterion, the superscript (1) indicating that one component was included in the regression. The weights were then recomputed for prediction from the two largest principal components

TABLE 19

Variable Parameters for Section 3.4

$n = 5, 10$
$\rho^2 = 0.25, 0.50, 0.75$
$\pi^2 = 0.20, 0.35, 0.50, 0.65, 0.80$
N = 20, 100 for n = 5 and N = 25, 105 for n = 10

resulting in cross-validities $r_c^{(2)}$. This procedure was continued until all components had been included in the regression. In general, the t largest principal components resulted in ten cross-validities $r_c^{(t)}$ for each t(t = 1, $\cdots$, n). The average of the *squares* of the ten cross-validities for each t was calculated. The largest of these averages is called $r_c^{2(max)}$ and occurs for $t = t_{max}$. The symbol t_{max} represents the number of components producing the largest average squared validity. (When $r_c^{(t)}$ was negative, a negative sign was affixed to its square before the averages were calculated.)

The above procedure was repeated for validating the principal components of $(C_2)_{xx}$ on C_1 . The averages of the squares of the ten cross-validities for each t as well as $r_c^{2(max)}$ and t_{max} were again calculated for each model.

In Section 3.3 it was shown that the squared cross-validity r_c^2 (when all variables or principal components are included in the regression) underestimates the squared population multiple correlation ρ^2. The result was confirmed with the 60 new models since only 14 out of the 120 values of $[r_c^{(n)}]^2$ exceeded ρ^2. Recall that $r_c^{(n)}$ is the same as r_c . The $r_c^{2(max)}$ were less biased as 30 out of 120 values exceeded ρ^2. Thus $r_c^{2(max)}$ is still an underestimate of the squared population multiple correlation.

Averaging the squared correlations *before* calculating t_{max} has the disadvantage of reducing $r_c^{2(max)}$ from what it would be if the maximum r_c^2 was found for each criterion and then these maxima averaged. These maxima will occur for different t for the different criteria, in general. In several cases which were examined, however, it was found that most maxima occurred for the same t. Since some averaging had to be done to comprehend the results, the method previously described was used for simplicity.

The results from these experiments will be displayed in two ways—first, by t_{max} (Table 20), and, second, by the average squared cross-validity (Figures 1 to 10). Table 20 shows how t_{max} varies as a function of ρ^2, π^2, and N for each of the two values of n. For example, the first section of Table 20 shows that, for all ten models with $\rho^2 = 0.25$ and $n = 5$ (two values of t_{max} per model), $t_{max} = 1$ occurred five times, $t_{max} = 2$ occurred four times, etc. This section of the table shows that t_{max} tended to increase as ρ^2 increased. This effect is stronger when $n = 10$. On the other hand, the effect of increasing π^2 is to decrease the value of t_{max} . Hence for larger values of π^2, a smaller number of principal components produce maximum cross-validities. This is true for both values of n. Finally, there is a tendency for larger values of N to lead to larger values of t_{max} . As N increases, more principal components are needed to maximize the cross-validity.

Figures 1 to 6 illustrate these results in a different way. In these figures, the average squared cross-validity is shown as a function of t, the number of principal components included in the regression. Each point in Figures 1 and 4 is the average of 12 cross-validities; each point in Figures 2 and 5 is the average of 20 cross-validities; and each point in Figures 3 and 6 is the

TABLE 20

Frequencies of t_{max} for Cross-Validities r_c^2

		n = 5 t_{max}							n = 10 t_{max}									
		1	2	3	4	5			1	2	3	4	5	6	7	8	9	10
ρ^2	.25	5	4	1	2	8	ρ^2	.25	16	1	0	0	0	0	0	0	1	2
	.50	6	1	0	3	10		.50	9	1	2	0	0	0	0	3	2	3
	.75	3	1	4	0	12		.75	3	2	1	0	0	1	2	0	2	9
π^2	.20	0	0	0	0	12	π^2	.20	2	0	1	0	0	0	0	2	2	5
	.35	0	2	2	2	6		.35	3	1	0	0	0	0	0	1	3	4
	.50	3	1	0	0	8		.50	8	0	0	0	0	1	0	0	0	3
	.65	4	1	3	2	2		.65	8	2	0	0	0	0	0	0	0	2
	.80	7	2	0	1	2		.80	7	1	2	0	0	0	2	0	0	0
N	20	10	4	3	4	9	N	25	17	3	3	0	0	1	0	3	2	1
	100	4	2	2	1	21		105	11	1	0	0	0	0	2	0	3	13

The cell entry is the frequency that the value of t_{max} occurred for models with the row value of the parameter.

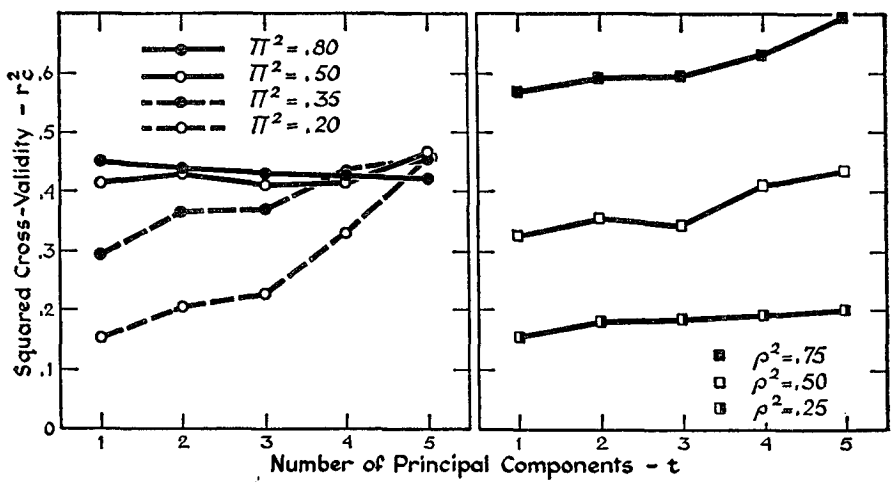

FIGURE 1 (Left)
Prediction from principal components, π^2 varying ($n = 5$)

FIGURE 2 (Right)
Prediction from principal components, ρ^2 varying ($n = 5$)

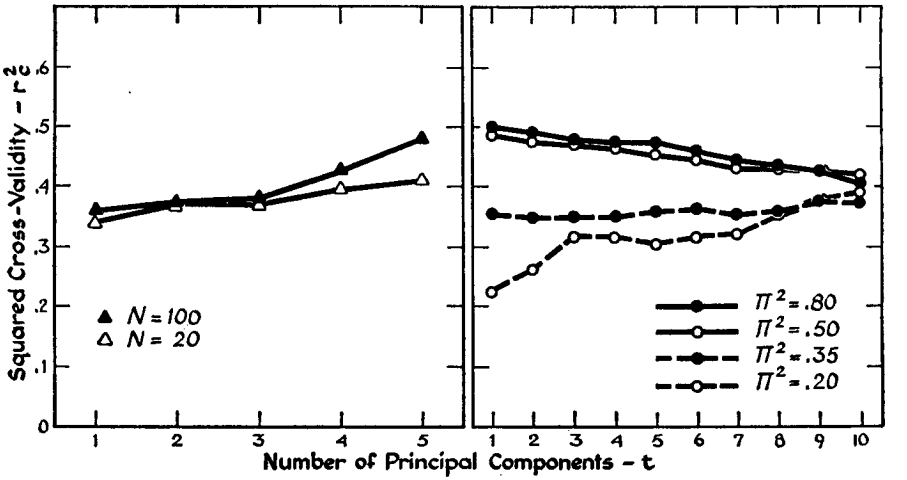

FIGURE 3 (Left)
Prediction from principal components, N varying (n = 5)

FIGURE 4 (Right)
Prediction from principal components, π^2 varying (n = 10)

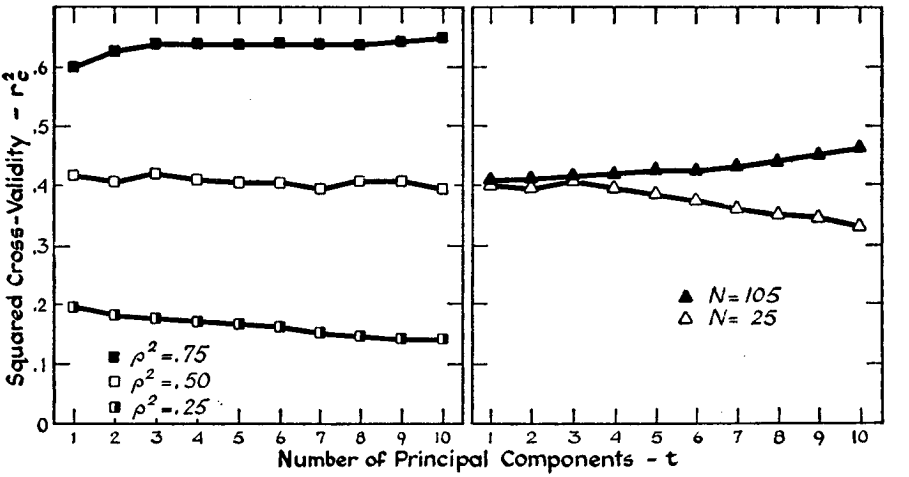

FIGURE 5 (Left)
Prediction from principal components, ρ^2 varying (n = 10)

FIGURE 6 (Right)
Prediction from principal components, N varying (n = 10)

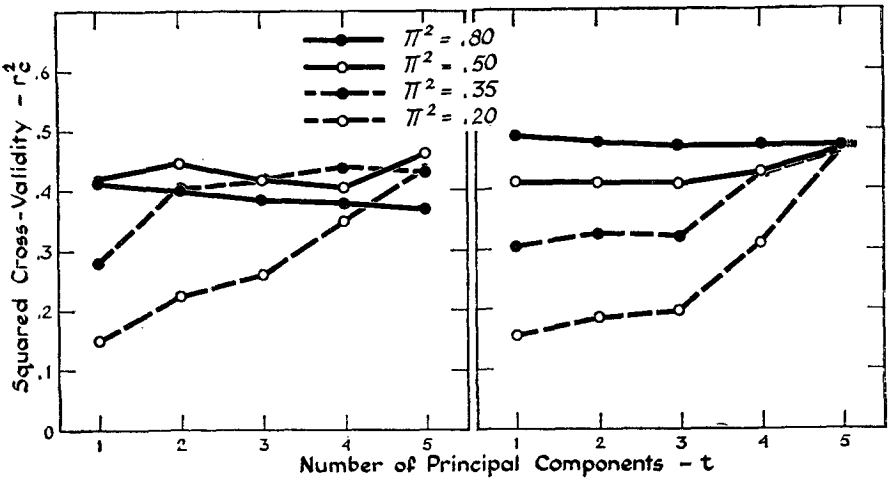

FIGURE 7 (Left)
Prediction from principal components, π^2 varying ($n = 5$, $N = 20$)

FIGURE 8 (Right)
Prediction from principal components, π^2 varying ($n = 5$, $N = 100$)

average of 30 cross-validities. (The results for $\pi^2 = 0.65$ have not been included in Figures 1 and 4 in order to avoid crowding the figures. In Figure 1, the $\pi^2 = 0.65$ curve would fall between the curves for $\pi^2 = 0.5$ and $\pi^2 = 0.8$. In Figure 4, the $\pi^2 = 0.65$ curve would fall slightly below the $\pi^2 = 0.5$ curve.)

Each figure shows that there is an interaction between the number of principal predictors (t) and the variable parameter (π^2, ρ^2, or N). This accounts for the results of Table 20. For example, when π^2 is small, an increase in t produces an increase in the average r_c^2. Hence for small π^2, the number of principal components producing maximum cross-validity is n or almost n. However, for large π^2, the average cross-validities are constant or slightly decreasing as a function of t. Therefore there is a tendency for one component, or at most a few components, to produce maximum cross-validity. In a similar way, the effects of ρ^2 and N can be compared in the figures and Table 20. The interaction of ρ^2 and t is quite small and will not be further discussed.

The present results may be compared with a previous major study of reduced rank prediction [Burket, 1964]. In Burket's Table 1, the prediction method 3 is prediction from the principal components of 29 predictors. In Burket's table it is clear that when the sample size is large (N = 255, 210, 165), the cross-validities decrease only very slowly as the number of principal components increase, while when the sample size is small (N = 75, 30), the cross-validities decrease significantly with t. While it is impossible to make a perfect comparison between Burket's results and the simulations, some similarity can be shown by considering Figures 7 to 10. These figures show

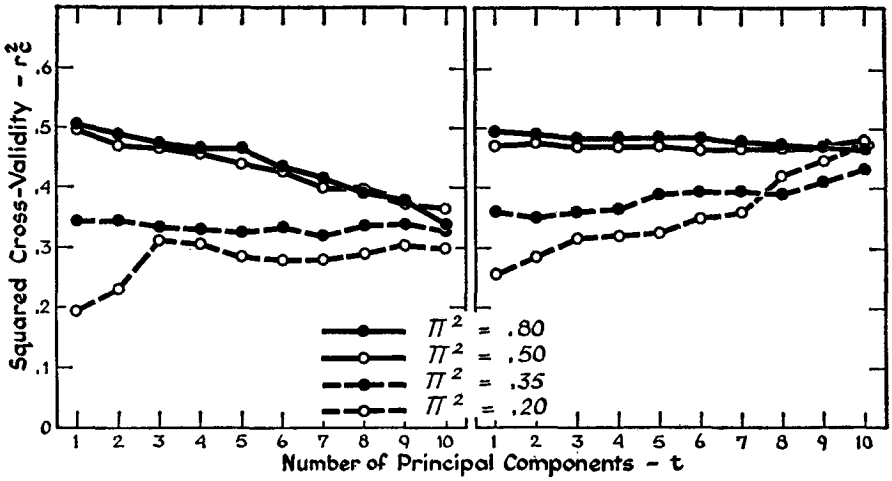

FIGURE 9 (Left)
Prediction from principal components, π^2 varying ($n = 10$, $N = 25$)

FIGURE 10 (Right)
Prediction from principal components, π^2 varying ($n = 10$, $N = 105$)

the simulation results for the two sample sizes separately instead of averaged together as in Figures 1 and 4. Now the experiments to be described later (Section 4.2) show that real data have large values of π^2 (approximately 0.8). It is reasonable to assume that Burket's data also have a large value of π^2 since the predictors and criteria in both studies are similar (Burket: predictors—Edwards Personal Preference Schedule, High School Grade-Point Averages, Test Scores, Age, Sex; criteria—Grade-Point Averages in Various College Course Areas; Herzberg: predictors—Sequential Tests of Educational Progress, School and College Ability Tests, Tests of General Interest; criteria—Scholastic Aptitude Tests, College Entrance Examination Board, Rank in High School Class). In Figures 7 to 10 it is seen that, for $\pi^2 = 0.8$, r_c^2 decreases only slightly when N is large but decreases much more rapidly when N is small, particularly when n = 10 (Figure 9). This corresponds to the results in Burket's Table 1.

Both Burket's study and the present one confirm that reduced rank prediction is of maximum benefit when the sample size is small. The present study also shows the great effect that the parameter π^2 has on prediction from the principal components. When π^2, the average criterion-related predictor variance, is small, r_c^2 increases rapidly with the number of principal components. However when π^2 is large, r_c^2 is constant or slightly decreasing as t increases. This result is easily understood, for, when π^2 is increased, one linear combination of the predictors, namely w_1, the first principal predictor, is increased in variance. The result is that the largest principal component

of the predictors becomes increasingly collinear with w_1 as π^2 increases. This means that a single principal component can produce better prediction than several components. Hence r_c^2 decreases slightly as t increases and therefore t_{max}, the value of t giving maximum r_c^2, tends to be near 1 as π^2 increases. (Note again that the results for $\pi^2 = 0.65$ have not been included in these figures. In Figures 7 and 8, the $\pi^2 = 0.65$ curve would almost coincide with the $\pi^2 = 0.8$ curve. In Figures 9 and 10 the $\pi^2 = 0.65$ curve would fall slightly below the $\pi^2 = 0.5$ curve.)

CHAPTER 4

STUDIES WITH SEVERAL CRITERIA

The generation of populations with several criterion variables permits the principal predictors to be used in multiple regression. In the first section of this chapter, simulation experiments with several criteria are described. The two reduced rank methods of prediction (multiple regression on the principal components and on the principal predictors) are compared. In the next section, some studies are described using real data from high school students. Finally, in the last section, an attempt is made to simulate the real data with the computer program.

4.1 Simulation Results

The purpose of this section is to compare the principal component and principal predictor methods of prediction in a number of models which differ in the distribution of (q_k^2, k = 1, $\cdots$, m) and π^2. The importance of the parameter π^2, the average criterion-related predictor variance, in prediction from the principal components of the predictors, has already been shown in the single criterion case (Section 3.4). When m = 1, there is only one q_k^2 and it is directly related to π^2($q_1^2 = n\pi^2$). However when there are several criteria, there are several q_k^2, each q_k^2 representing the total dependence of the predictors on the kth principal predictor. The sum of the q_k^2 is equal to $n\pi^2$.

Similarly, the (D_{kk}^2, k = 1, $\cdots$, m) are the total dependence of the criteria on each principal predictor. Since the D_{kk}^2 are eigenvalues (of $\Sigma_{\hat{y}\hat{y}}$), they are always considered in descending order. In this section, the D_{kk}^2 distribution was kept constant. By varying the order of the q_k^2 it is possible to vary the relative dependence of the predictors on the principal predictors. This variation will produce differences in the effectiveness of the two methods of prediction.

Eighteen models were studied. The constant input parameters are shown in Table 21 and the variable parameters in Table 22. For each value of π^2, three different q_k^2 distributions were used, called decreasing, level, and increasing. The same distributions, except for scaling by π^2, were used for all three values of π^2. Two models, differing only in sample size, were generated for each combination of π^2 and q_k^2 distribution. A pair of samples of size 20 (small N) was generated from one model and a pair of samples of size 75 (large N) was generated from the second model.

TABLE 21

Constant Parameters for Section 4.1

$n = 10$

$m = 5$

$n_s = 10$

$\rho^2 = 0.6$

$D_{11}^2 = 1.2$, $D_{22}^2 = 0.8$, $D_{33}^2 = 0.6$,
$D_{44}^2 = 0.3$, $D_{55}^2 = 0.1$

$v_x = 0.01$

$e_x = 0.005$

$v_y = 0.01$

$e_y = 0.005$

number of Σs per model = 1

number of C_1, C_2 pairs per Σ = 1

In the *decreasing* q_k^2 distribution, half the dependence of the predictors on the principal predictors is dependence on the first principal predictor. When $\pi^2 = 0.8$, this means that 40% of the total variance of the predictors is linearly dependent on the first principal predictor. The contribution of the succeeding principal predictors is progressively smaller. When $\pi^2 = 0.2$ and the q_k^2 are again decreasing, the first principal predictor accounts for only 10% of the predictor variance and the other principal predictors account for less, with a total of 20% of the variance of the predictors explained by the principal predictors.

In the *level* q_k^2 distribution, each principal predictor contributes equally to the predictors, the contribution of each varying from 16% when $\pi^2 = 0.8$ to 4% when $\pi^2 = 0.2$.

When the (q_k^2, k = 1, $\cdots$, m) are *increasing* the dependence is exactly reversed from the decreasing case. Most of the dependence of the predictors on the principal predictors is dependence on the fifth (last) principal predictor. The dependence on the first principal predictor is very small.

One change was made in the generation of samples for the calculations in this chapter. The sample covariance matrices, C_1 and C_2, were changed to *correlation* matrices in order to make possible the calculation of a sample π^2 and sample (q_k^2, k = 1, $\cdots$, m).

The calculations performed on the sample correlation matrices were similar to the calculations in Section 3.4 except that two methods of pre-

diction were compared and the criteria were no longer duplicate criteria. Multiple regression on the principal components of the predictors was done first. The correlation matrix of the predictors, $(C_1)_{xx}$, was diagonalized and the weights for predicting each of the five criteria from the largest component were calculated. The cross-validities for each criterion in the second sample were calculated. The average, over criteria, of the squares of these validities was calculated and is presented in Tables 23 to 28 in the "r_c^2" columns under "P. C." for $t = 1$ (one component). The weights were then recomputed for the two largest components ($t = 2$) and the average squared cross-validity calculated. This was repeated for $t = 3, \cdots, 10$. The whole procedure was repeated again for validating the weights derived in C_2 on C_1, but these results are not reproduced in the tables as they are very similar to the validation of C_1 on C_2.

Prediction from the principal predictors was then performed following the method given in Section 2.7. The weights in (2.7.11) were calculated in the first sample for each value of t ($t = 1, \cdots, 5$) and the weights were cross-validated in the second sample. The average of the cross-validities for each value of t was calculated and is given in Tables 23 to 28 in the "r_c^2" columns under "P. P." for each t. The validation of sample 2 on sample 1 is not reported here.

In the derivation sample the following sample statistics were calculated: sample (q_k^2, $k = 1, \cdots, 5$), sample π^2, sample (D_{kk}^2, $k = 1, \cdots, 5$), and sample ρ^2. For the calculation of the sample (q_k^2, $k = 1, \cdots, 5$) and the sample π^2, see (2.7.15) and (2.7.16). The D_{kk}^2 are the eigenvalues of $C_{\phi\phi}$ and the sample ρ^2 is r^2, the average squared multiple correlation using all predictors. The sample ρ^2 is the average of the (D_{kk}^2, $k = 1, \cdots, 5$).

TABLE 22

Variable Parameters for Section 4.1

$\pi^2 = 0.2, 0.5, 0.8$

q_k^2 distribution--decreasing, level, increasing as shown below

$N = 20, 75$

	decreasing	level	increasing
q_1^2	$5.0\ \pi^2$	$2.0\ \pi^2$	$0.625\ \pi^2$
q_2^2	$2.5\ \pi^2$	$2.0\ \pi^2$	$0.625\ \pi^2$
q_3^2	$1.25\ \pi^2$	$2.0\ \pi^2$	$1.25\ \pi^2$
q_4^2	$0.625\ \pi^2$	$2.0\ \pi^2$	$2.5\ \pi^2$
q_5^2	$0.625\ \pi^2$	$2.0\ \pi^2$	$5.0\ \pi^2$

TABLE 23

Several Criteria (N = 20) q_k^2 Distribution Decreasing

t or k	$\pi^2 = .2$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	q_k^2	$\pi^2 = .5$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	q_k^2	$\pi^2 = .8$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	q_k^2
1	.06	.21	1.5	.6	-.01	-.01	2.3	2.6	.28	.19	1.8	2.6
2	.09	.13	1.0	.3	.23	.10	1.2	.7	.31	.19	.9	2.3
3	.09	.23	.9	1.0	.41	-.01	.3	.4	.33	.20	.8	1.4
4	.11	.33	.5	.4	.41	-.01	.2	.6	.25	.23	.2	1.0
5	.17	.30	.2	.4	.32	-.01	.0	1.5	.29	.23	.2	.3
6	.22				.43				.30			
7	.41				.45				.34			
8	.44				.47				.25			
9	.44				.49				.23			
10	.30				-.02				.23			
Sample ρ^2			.80				.80				.77	
Sample π^2				.27				.57				.75

TABLE 24

Several Criteria (N = 20) q_k^2 Distribution Level

t or k	$\pi^2 = .2$ P.C r_c^2	P.P. r_c^2	Sample D_{kk}^2	q_k^2	$\pi^2 = .5$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	q_k^2	$\pi^2 = .8$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	q_k^2
1	-.00	.10	1.7	.2	.03	-.01	2.0	.8	.11	-.00	1.9	1.2
2	.02	.12	1.4	.9	.08	.08	1.3	1.2	.21	.14	.8	.9
3	.05	.12	.5	.5	.15	.12	.7	1.1	.20	.26	.6	2.1
4	.04	.12	.3	.8	.14	.14	.2	1.3	.22	.29	.1	1.4
5	.04	.13	.2	1.4	.21	.15	.1	1.4	.16	.30	.0	.9
6	.10				.23				.21			
7	.14				.22				.22			
8	.18				.28				.33			
9	.20				.31				.32			
10	.13				.15				.30			
Sample ρ^2			.83				.87				.69	
Sample π^2				.37				.58				.66

TABLE 25

Several Criteria (N = 20) q_k^2 Distribution Increasing

t or k	$\pi^2 = .2$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	q_k^2	$\pi^2 = .5$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	q_k^2	$\pi^2 = .8$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	q_k^2
1	.03	.04	1.6	.4	.11	.18	1.5	.6	.04	.12	1.4	.7
2	.02	.04	.9	.9	.17	.33	1.1	1.1	.11	.13	1.0	1.0
3	.07	.04	.8	.5	.19	.29	.7	1.7	.18	.22	.8	1.1
4	.02	.04	.6	.6	.20	.43	.5	1.7	.20	.26	.4	1.0
5	.02	.04	.1	1.7	.14	.44	.3	.5	.30	.28	.1	4.1
6	.05				.13				.31			
7	.10				.25				.30			
8	.23				.38				.29			
9	.27				.47				.38			
10	.04				.44				.25			
Sample ρ^2			.80				.80				.75	
Sample π^2				.42				.56				.78

TABLE 26

Several Criteria (N = 75) q_k^2 Distribution Decreasing

t or k	$\pi^2 = .2$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2	$\pi^2 = .5$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2	$\pi^2 = .8$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2
1	-.00	.15	1.5	1.1	.19	.27	1.2	2.1	.22	.23	1.3	4.1
2	.07	.21	.9	.7	.26	.42	.9	1.7	.38	.34	.6	1.7
3	.07	.27	.6	.3	.27	.53	.6	.5	.46	.44	.5	1.4
4	.07	.32	.2	.3	.37	.54	.3	.3	.48	.45	.4	.6
5	.15	.33	.1	1.4	.40	.56	.1	.8	.51	.45	.1	.7
6	.26				.44				.50			
7	.29				.44				.51			
8	.37				.52				.53			
9	.44				.54				.53			
10	.33				.56				.45			
Sample ρ^2			.65				.63				.58	
Sample π^2				.38				.54				.85

TABLE 27

Several Criteria (N = 75) q_k^2 Distribution Level

t or k	$\pi^2 = .2$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2	$\pi^2 = .5$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2	$\pi^2 = .8$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2
1	.02	.12	1.3	.5	.00	.21	1.3	1.0	.09	.21	1.3	1.7
2	-.02	.26	.9	.5	.06	.33	.7	.7	.19	.33	1.0	1.7
3	-.01	.32	.7	.7	.10	.41	.6	1.1	.33	.39	.6	1.9
4	.03	.37	.4	.4	.11	.46	.3	.9	.46	.45	.2	1.2
5	.07	.38	.1	.5	.22	.48	.2	.9	.49	.45	.1	1.5
6	.12				.34				.51			
7	.22				.40				.52			
8	.34				.43				.52			
9	.33				.48				.52			
10	.38				.48				.45			
Sample ρ^2			.69				.63				.64	
Sample π^2				.26				.46				.82

TABLE 28

Several Criteria (N = 75) q_k^2 Distribution Increasing

t or k	$\pi^2 = .2$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2	$\pi^2 = .5$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2	$\pi^2 = .8$ P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2
1	-.01	.17	1.2	.2	-.00	.26	1.4	.4	.02	.23	1.6	.8
2	.00	.20	1.0	.7	.04	.33	.8	.5	.06	.39	1.0	.8
3	.01	.27	.5	.8	.08	.43	.7	.7	.17	.46	.6	1.3
4	.01	.34	.4	1.1	.10	.46	.3	1.7	.22	.51	.2	1.4
5	.05	.36	.1	.7	.16	.48	.1	2.2	.47	.51	.1	2.6
6	.08				.21				.49			
7	.13				.28				.53			
8	.26				.41				.55			
9	.39				.46				.53			
10	.36				.48				.51			
Sample ρ^2			.65				.67				.70	
Sample π^2				.35				.56				.68

Consider first N = 75 (Tables 26 to 28). When $\pi^2 = 0.2$, the first few principal predictors are far superior to the first few principal components in average cross-validity. This is true whether the q_k^2 distribution is decreasing, level, or increasing. The reason is that the principal predictors account for only 20% of the predictor variance when $\pi^2 = 0.2$. The largest principal components therefore reflect variance mostly independent of the principal predictors and therefore the largest principal components are not good predictors. The principal predictors, though of small variance, are good predictors and multiple regression on them cross-validates well.

The situation changes, however, as π^2 increases to 0.5 and 0.8. When $\pi^2 = 0.8$ and the q_k^2 distribution is decreasing or level (*not* increasing), there is practically no difference between the two prediction methods in the average cross-validity for $t = 1, \cdots, 5$. When $\pi^2 = 0.8$, the five principal predictors account for a total of 80% of the predictor variance and therefore the principal components of the predictors are very similar to the principal predictors. Therefore the principal components and principal predictors cross-validate equally well.

An exception occurs for the *increasing* q_k^2 distribution when $\pi^2 = 0.8$ (Table 28). Here the principal predictors cross-validate much better than the first few principal components. The reason is that the *first* principal predictor (largest D_{kk}^2 and hence best predictor) is the smallest principal predictor in terms of associated predictor variance (q_1^2 is the smallest q_k^2). The principal predictor method of regression properly picks this first principal predictor as the best predictor. The principal component method of prediction, however, chooses the principal predictors in reverse order, since the principal predictor with largest predictor variance (40%) is the fifth principal predictor and the fourth principal predictor has the next largest variance (20%), etc. Note that there is a large increase in average cross-validity between $t = 4$ and $t = 5$ principal components in this case. The fifth principal component is approximately collinear with the first principal predictor which is the best predictor. On the other hand, when $\pi^2 = 0.2$, the large increase in average cross-validity using principal components does not occur until t = 8 or 9.

The results when $\pi^2 = 0.5$ are intermediate between those for $\pi^2 = 0.2$ and $\pi^2 = 0.8$. Furthermore, the results for N = 20 (Tables 23 to 25) are similar to the N = 75 results just described except that the effects are not as clear due to instability of the weights with small N. An interesting effect is shown in two cases, however (Table 23, $\pi^2 = 0.5$ and Table 25, $\pi^2 = 0.2$). This effect is not dependent on sample size and could have occurred for N = 75. In both cases the average cross-validity when all factors are included in the regression is essentially zero, since the predictors are dependent. The population Σ_{xx} which was generated in these two cases was almost singular. This was shown by the difficulty in inverting it. Every time a

matrix is inverted in the simulation program, the result is checked by multiplying the inverse by the original matrix and comparing the result with the identity matrix. The largest difference, in absolute value, between corresponding elements in the two matrices is printed as WINV. Most generated Σ_{xx} matrices yield WINV = 0.00001 or less. In the two cases mentioned above, WINV = 0.003 and 0.0001, respectively, indicating approximate dependence of the predictors in the population. This dependence appears in the generated samples as well.

In the two singular cases, prediction from the principal components, for $t < 10$, is successful. However, for all values of t, the cross-validities using the principal predictors as predictors are practically zero. The t largest principal components ($t < 10$) are independent and their weights cross-validate well. This is an advantage of prediction from the principal components of the predictors—the effect of dependence of the predictors can be eliminated. However, the weights on the principal predictors are not stable in the singular case, regardless of the number of principal predictors included in the regression.

Even though variation of the population parameters has an appreciable effect on prediction by the two reduced rank methods, for practical application of these results it would be necessary to determine from *samples* what the population parameters are. Can these parameters be estimated? The sample (q_k^2, $k = 1, \cdots, 5$) and π^2 are estimates of the corresponding population parameters and are shown in Tables 23 to 28. In general the sample q_k^2 distribution is similar to the population distribution. This is shown most clearly for large sample size (N = 75). When the population q_k^2 distribution is decreasing, the largest sample q_k^2 generally occurs for $k = 1$. When the q_k^2 distribution is level, the sample values, q_k^2, are approximately equal. When the q_k^2 distribution is increasing, the largest sample q_k^2 normally occurs for $k = 5$. It is therefore possible to decide, on the basis of the (q_k^2, $k = 1, \cdots, m$) in the *derivation* sample, whether the principal predictors cross-validate better than the principal components or whether there will be little difference between the two methods.

As an additional aid in making this determination, it is important to estimate π^2. This may be done from the value of π^2 computed in the derivation sample. It can be seen from the tables that the sample π^2 is a rough measure of the population value; there is a tendency for the sample value to shift nearer 0.5 than the population value.

Since the population (D_{kk}^2, $k = 1, \cdots, 5$) were not varied in these simulation studies, the sample values, D_{kk}^2, shown in the tables are relatively constant from sample to sample. These parameters will be discussed further in the next section.

This section has shown that prediction from the principal predictors is an effective method of prediction, particularly when the q_k^2 distribution is

increasing or π^2 is small. In other cases prediction from the principal components is almost as successful as prediction from the principal predictors. The only case in which prediction from the principal components is superior to prediction from the principal predictors is when the predictors are dependent.

4.2 Study of Real Data

The calculations described in the preceding section were also performed on some samples of real data. The data were collected in 1961 by the Educational Testing Service, Princeton, N. J., from 1205 boys in academic high schools. The 21 variables employed in the multiple regression calculations are listed in Table 29. There are two sets of eight predictors each and one set of five criterion variables. The first set of predictors, called the S-predictors, consists of six variables from the Sequential Tests of Educational Progress (STEP) and two variables from the School and College Ability Tests (SCAT). The second set of predictors, called the T-predictors, consists of eight variables from the Tests of General Interest (TGI). The criterion variables are two variables from the Scholastic Aptitude Test (SAT), two variables from the College Entrance Examination Board (CEEB), and the rank in the high school class.

Eight samples were drawn at random from the pool of 1205 subjects. Four of the samples were of size N = 20 (Samples 1, 2, 3, 4) and the other four samples were of size N = 75 (Samples 5, 6, 7, 8). Four double cross-validations were performed (two for each sample size) using the S-predictors. Then the same samples were used in four double cross-validations using the T-predictors.

Tables 30 (N = 20) and 31 (N = 75) are a summary of the calculations made on these samples; the calculations were the same as those made on the simulation samples in Section 4.1. Correlation matrices were used. Again, only the validation of C_1 on C_2 is reported.

In most of the samples, the principal predictors (P. P.) validate more poorly than the principal components (P. C.), and in the two cases where the first principal predictor validates better than the first principal component, the improvement is not great. Another feature of these data is the nearly constant average cross-validity of the principal predictors for t = 1, $\cdots$, 5. Even though, in some cases, a few principal *components* are far superior to including all predictors in the regression, in no case is the first principal *predictor* significantly better than all predictors.

These findings can be understood by considering the estimates of the parameters ρ^2, π^2, (D^2_{kk} , k = 1, $\cdots$, 5), and (q^2_k , k = 1, $\cdots$, 5). In all four derivation samples, the sample π^2 is at least 0.87 for the S-predictors and at least 0.78 for the T-predictors. Therefore these samples correspond approximately to the π^2 = 0.8 cases of Section 4.1. Furthermore the sample q^2_k

TABLE 29

E. T. S. Variables

S-predictors
STEP Mathematics
STEP Science
STEP Social Studies
STEP Reading
STEP Listening
STEP Writing
SCAT Verbal
SCAT Quantitative

T-predictors
TGI Industrial Arts
TGI Home Arts
TGI Physical Education
TGI Biological Science
TGI Music and Art
TGI History-Literature
TGI Entertainment
TGI Public Affairs

Criteria
SAT Verbal
SAT Mathematical
CEEB English Composition
CEEB American History
Rank in High School Class

distributions are in all cases but one heavily weighted on q_1^2, thus indicating an extremely *decreasing* q_k^2 distribution. Returning to the corresponding simulation examples in Section 4.1, it is seen that the ETS results do not differ greatly from the last columns of Tables 23 (N = 20) and 26 (N = 75).

The failure of the principal predictors in the ETS samples can be further explained by the sample D_{kk}^2 distribution. In all cases, D_{11}^2 is at

TABLE 30

E. T. S. Samples (N = 20)

t or k	Sample 1 validated on Sample 2				Sample 3 validated on Sample 4			
	P.C.	P.P.	Sample		P.C.	P.P.	Sample	
	r_c^2	r_c^2	D_{kk}^2	q_k^2	r_c^2	r_c^2	D_{kk}^2	q_k^2
			S-predictors					
1	.56	.41	3.15	4.68	.47	.42	3.93	5.83
2	.55	.42	.40	.83	.47	.41	.28	.34
3	.56	.46	.24	.59	.42	.39	.11	.35
4	.53	.44	.11	.41	.44	.36	.03	.43
5	.52	.44	.02	.46	.45	.37	.01	.26
6	.52				.44			
7	.51				.40			
8	.44				.37			
Sample ρ^2			.78				.87	
Sample π^2				.87				.90
			T-predictors					
1	.26	.09	2.16	1.69	.28	.35	3.51	4.31
2	.27	.06	.27	.68	.31	.32	.25	.42
3	.27	.08	.16	1.93	.33	.32	.09	.47
4	.27	.07	.06	1.57	.32	.35	.04	.68
5	.15	.08	.02	.37	.32	.34	.01	.51
6	.13				.30			
7	.12				.32			
8	.08				.34			
Sample ρ^2			.54				.78	
Sample π^2				.78				.80

least 80% of the total predictable variance and therefore the D_{kk}^2 distribution in the ETS data is weighted more in favor of D_{11}^2 than the populations considered in Section 4.1. This means that the first principal component is very similar to the first principal predictor and hence prediction from the principal components is effective.

In the next section some simulation models will be considered that more closely match the ETS data, particularly in the D_{kk}^2 distribution.

4.3 Simulation of Real Data

The ETS data samples differ from the simulated data in Section 4.1 in several respects. The (D_{kk}^2, k = 1, $\cdots$, 5) distribution is much more con-

TABLE 31

E. T. S. Samples (N = 75)

t or k	Sample 5 validated on Sample 6				Sample 7 validated on Sample 8			
	P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2	P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2
				S-predictors				
1	.65	.61	3.28	5.55	.64	.66	3.29	5.87
2	.66	.63	.13	.57	.66	.69	.09	.47
3	.66	.63	.07	.40	.67	.69	.06	.27
4	.64	.63	.02	.27	.67	.69	.03	.36
5	.64	.63	.01	.24	.67	.69	.01	.22
6	.64				.68			
7	.63				.68			
8	.63				.69			
Sample ρ^2			.70				.70	
Sample π^2				.88				.90
				T-predictors				
1	.49	.37	2.47	3.54	.48	.46	2.09	4.15
2	.49	.37	.04	.82	.47	.45	.09	.53
3	.48	.37	.02	.62	.45	.44	.04	.51
4	.44	.37	.01	.64	.46	.43	.02	.76
5	.44	.37	.00	.62	.46	.43	.01	.67
6	.40				.45			
7	.39				.44			
8	.37				.43			
Sample ρ^2			.51				.45	
Sample π^2				.78				.83

centrated on D_{i1}^2 in the ETS data than in Section 4.1 where the population D_{kk}^2 distribution was fixed as (1.2, 0.8, 0.6, 0.3, 0.1). The distribution of $(q_k^2, k = 1, \cdots, 5)$ in the ETS data is similar to the decreasing q_k^2 distribution in Section 4.1 although, in most cases, the ETS data had an even larger q_1^2.

The S-predictors were superior to the T-predictors; the approximate population ρ^2 for the S-predictors might be estimated to be 0.65 and for the T-predictors about 0.45. Other parameters were roughly estimated for the two sets of predictors and are shown in Table 32. These parameters

TABLE 32

Estimated Parameters Used to Simulate E. T. S. Data

$n = 8$

$m = 5$

$n_s = 8$

$v_x = 0.01$

$e_x = 0.005$

$v_y = 0.01$

$e_y = 0.005$

$N = 75$

number of Σs per predictor type = 2

number of C_1, C_2 pairs per $\Sigma = 1$

S-predictors (First Σ and Second Σ)

$\rho^2 = 0.65$

$D_{11}^2 = 3.0$, $D_{22}^2 = 0.1$, $D_{33}^2 = 0.05$,
$D_{44}^2 = 0.05$, $D_{55}^2 = 0.05$

$\pi^2 = 0.85$

$q_1^2 = 5.0$, $q_2^2 = 0.6$, $q_3^2 = 0.4$, $q_4^2 = 0.4$, $q_5^2 = 0.4$

T-predictors (Third Σ and Fourth Σ)

$\rho^2 = 0.45$

$D_{11}^2 = 2.0$, $D_{22}^2 = 0.1$, $D_{33}^2 = 0.05$,
$D_{44}^2 = 0.05$, $D_{55}^2 = 0.05$

$\pi^2 = 0.75$

$q_1^2 = 4.5$, $q_2^2 = 0.5$, $q_3^2 = 0.4$, $q_4^2 = 0.3$, $q_5^2 = 0.3$

TABLE 33

Simulation of E. T. S. Samples (N = 75)

S-predictors

t or k	First Σ Sample A validated on Sample B				Second Σ Sample C validated on Sample D			
	P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2	P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2
1	.56	.57	3.24	5.05	.59	.57	3.27	5.24
2	.58	.58	.15	.44	.59	.60	.16	.56
3	.60	.60	.12	.38	.61	.59	.11	.32
4	.60	.61	.03	.45	.63	.60	.05	.41
5	.61	.62	.02	.56	.64	.61	.01	.55
6	.60				.61			
7	.61				.61			
8	.62				.61			
Sample ρ^2			.71				.72	
Sample π^2				.86				.89

T-predictors

t or k	Third Σ Sample E validated on Sample F				Fourth Σ Sample G validated on Sample H			
	P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2	P.C. r_c^2	P.P. r_c^2	Sample D_{kk}^2	Sample q_k^2
1	.39	.41	2.04	4.21	.51	.54	1.67	4.14
2	.39	.37	.34	.27	.51	.55	.20	.41
3	.40	.36	.19	.51	.52	.53	.13	.86
4	.40	.38	.12	.51	.52	.52	.09	.28
5	.41	.39	.03	.57	.51	.52	.03	.35
6	.42				.52			
7	.44				.53			
8	.39				.52			
Sample ρ^2			.54				.42	
Sample π^2				.76				.75

were used to generate population and sample correlation matrices using the simulation program. Two populations (First Σ and Second Σ) were generated for the S-predictor simulation and two populations (Third Σ and Fourth Σ) were generated for the T-predictor simulation. One pair of sample correlation matrices was generated for each population ($N = 75$ in all cases). The results of the validation of C_1 on C_2 for each of the four populations are shown in Table 33.

If these tables are compared with the corresponding Table 31 for the real data, it will be seen that the real and simulation results correspond closely. By adjusting the parameters it would be possible to make the match even better, but the simulation using the parameters in Table 32 is presented here, since this simulation was the first attempted. The close similarity of the simulation results to the results from the ETS data shows that the simulation model is basically sound.

CHAPTER 5

SUMMARY AND CONCLUSIONS

The several methods of multiple regression discussed in this study are designed to provide optimal weights for predictor variables. The weights are optimal in the sense that, in new samples, the weighted linear combination of the predictors has the highest possible correlation with the criterion variable. By means of cross-validation, it is possible to estimate the correlation in new samples using only data in a (divided) original sample.

For any given problem it is important to decide which prediction method gives the best weights. If one exhaustively tries all prediction methods it is straightforward, using cross-validation, to pick the best linear combination of the predictors. But there are some disadvantages to this procedure. It is lengthy even with a computer; there is capitalization on chance results; and the procedure does not provide a way to generalize to new variables or new populations.

It is apparent that no one method of prediction will be optimal for all possible predictor and criterion distributions. Even if one method, for example prediction from the principal components, were superior, it would still be necessary to decide the number of components to include in the regression. Burket's [1964] work included the computation of statistics which were of some assistance in deciding how many principal components to include in the regression. The present study considered some fundamental parameters of the population distribution which are relevant to the choice of prediction method and the number of components to include in the regression.

In order to study the effect of these parameters on prediction, the distributions were simulated on a computer. The parameters were systematically varied and the prediction methods were compared for each parameter set by applying the weights to cross-validation samples.

In Section 3.3 the accuracy of the sample multiple correlation and cross-validity as measures of the population multiple correlation and cross-validity were studied. The squared sample multiple correlation, r^2, is an over-estimate of the squared population multiple correlation, ρ^2. The bias tends to decrease with increasing sample size and to increase with increasing number of predictors and increasing ρ^2. The bias is correctly estimated by formulas of Wishart [1931] and Wherry [1931]. The sample and population cross-validities are approximately equal and underestimate ρ^2. The sample

cross-validity is therefore a good estimate of the population cross-validity but not of the population multiple correlation.

The dependence of the cross-validation of the principal components of the predictors on the distribution parameters was considered in Section 3.4. The technique used was to calculate the number of principal components which produced maximum cross-validity. This number, called t_{max}, was studied as a function of four parameters—the sample size, N, the number of predictors, n, the squared population multiple correlation, ρ^2, and the average criterion-related predictor variance, π^2. It was found that t_{max} is an increasing function of N and ρ^2 and a decreasing function of n and π^2. This means that a few (1 or 2, say) principal components will be more effective than many components when N is small, n is large, ρ^2 is small, and π^2 is large.

Many prediction problems in psychology involve multiple criteria, no one of which can be considered to be *the* criterion. A convenient way to avoid choosing one criterion, and at the same time, achieve some synthesis of the criteria, is to weight each standardized criterion equally and to optimize the prediction of all criteria simultaneously. The effectiveness of any prediction method can then be estimated from the average squared cross-validity.

Two prediction methods were compared in this way. The first, prediction from the largest principal components of the predictors, does not use criterion information in the selection of the components and may be used for one or several criteria. The second, prediction from the principal predictors, uses criterion information to calculate the principal predictors themselves. This method optimizes the average squared multiple correlation in the derivation sample.

It was found, for the distributions studied, that the principal predictors had superior or equal cross-validities to the principal components except when the predictors were approximately dependent. The superiority of principal predictors was particularly evident when π^2 was small and the q_k^2 distribution was increasing, meaning that the first principal predictor accounted for much less of the predictor variance than the last principal predictor. However this combination of parameters—π^2 small and the q_k^2 distribution increasing—may occur rarely, if at all, in real multivariate distributions. In the sample of real ability and interest data from the Educational Testing Service, π^2 was very large and the q_k^2 distribution was decreasing with heavy concentration on q_1^2. In these data, as in the corresponding simulation data, the principal components were superior to the principal predictors.

This result is similar to Burket's [1964] finding that the principal components correlating greatest with the criterion do not validate as well as the largest principal components. It appears to be an advantage to select linear combinations of the predictors independently of criterion information in order to maximize cross-validity.

In order for the conclusions of a simulation study to apply to real pre-

diction situations, it must be shown that the simulated distributions are similar to real distributions in relevant characteristics. Several sections of this study were concerned with this demonstration. In Section 3.1 it was shown that, when the population multiple correlation is zero, the simulation sample statistics (multiple correlation and cross-validity) obey known statistical laws. In Section 3.2 it was shown that changes in the covariance matrix of the predictors, keeping the multiple correlation constant, have no effect on the correlation statistics. Finally, in Section 4.3, several models and samples were generated in order to match the ETS data more closely. The results from this simulation were almost identical to the ETS results.

It would be interesting to determine if other real data have different values of π^2 and different q_k^2 distributions from those of the ETS data and to see if calculations using these variables obey the laws discovered in simulation. It is also necessary to extend the calculations to larger numbers of predictors and criteria. Such work would be a further check on the effectiveness of the simulation model which was used in this study.

REFERENCES

Anderson, T. W. *An introduction to multivariate statistical analysis.* New York: Wiley, 1958.

Bartlett, M. S. On the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh,* 1933, **53,** 260–283.

Box, G. E. P., & Muller, M. E. A note on the generation of random normal deviates. *The Annals of Mathematical Statistics,* 1958, **29,** 610–611.

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

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

Efroymson, M. A. Multiple regression analysis. In A. Ralston & H. S. Wilf (Eds.), *Mathematical methods for digital computers.* New York: Wiley, 1960. Pp. 191–203.

Ezekiel, M., & Fox, K. A. *Methods of correlation and regression analysis.* (3rd ed.) New York: Wiley, 1959.

Fisher, R. A. The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society,* 1928, A, **121,** 654–673.

Guttman, L. To what extent can communalities reduce rank? *Psychometrika,* 1958, **23,** 297–308.

Hase, H. D., & Goldberg, L. R. Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin,* 1967, **67,** 231–248.

Herzberg, P. A. The parameters of cross-validation. Unpublished doctoral dissertation, Univ. of Illinois, Urbana, 1967.

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

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

Hotelling, H. The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology,* 1957, **10,** 69–79.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics.* Vol. 2. London: Griffin, 1961.

Kshirsagar, A. M. Bartlett decomposition and Wishart distribution. *The Annals of Mathematical Statistics,* 1959, **30,** 239–241.

Larson, S. C. The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology,* 1931, **22,** 45–55.

Leiman, J. M. The calculation of regression weights from common factor loadings. Unpublished doctoral dissertation, Univ. of Washington, 1951.

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

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

Mosier, C. I. Problems and designs of cross-validation. *Educational and Psychological Measurement,* 1951, **11,** 5–11.

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

Olkin, I., & Pratt, J. W. Unbiased estimation of certain correlation coefficients. *The Annals of Mathematical Statistics,* 1958, **29,** 201–211.

Owen, D. B. *Handbook of statistical tables.* Reading, Mass.: Addison-Wesley, 1962.

Siegel, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

Tucker, L. R. Transformation of predictor variables to a simplified regression structure. Unpublished report. Princeton, N. J.: Educational Testing Service, 1957.

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

Wijsman, R. A. Random orthogonal transformations. *The Annals of Mathematical Statistics*, 1957, **28**, 415–423.

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