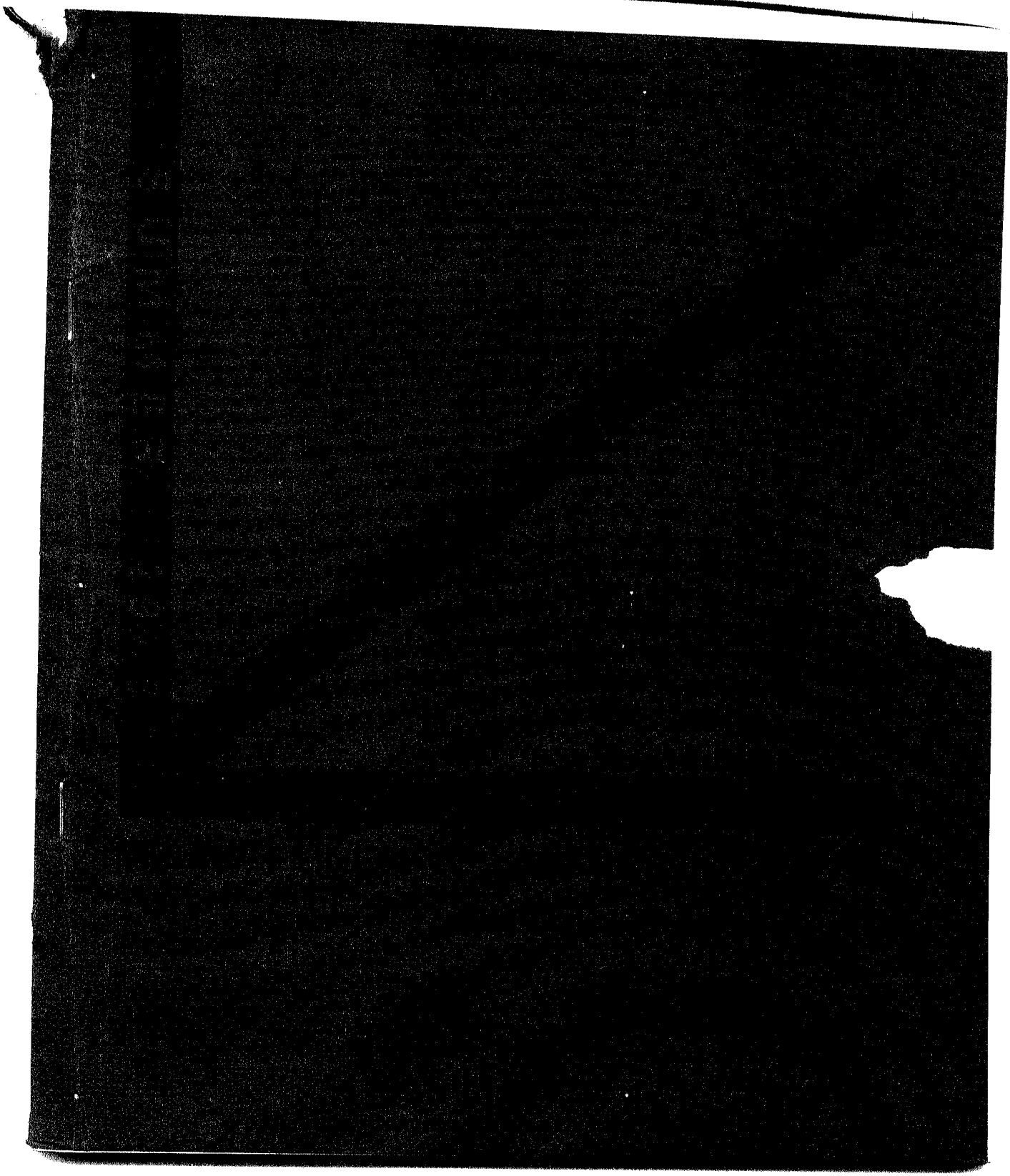


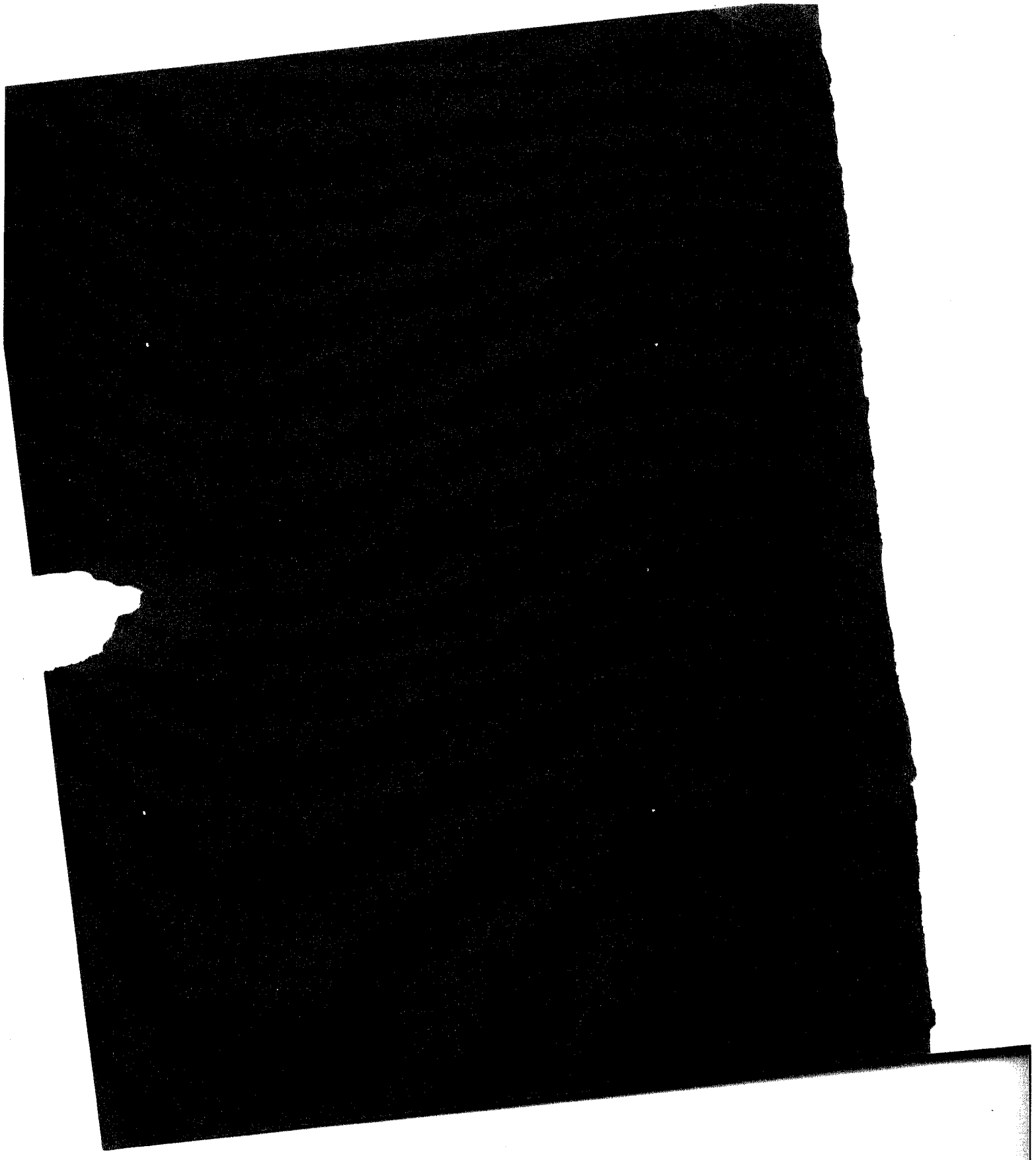


TABLE OF CONTENTS

MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 15, NUMBER 1, SUMMER 1986

Catalytic Variables for Improving Personnel Classification and Assignment

Joe H. Ward, Jr. and Richard C. Sorenson
Navy Personnel Research and Development Center
San Diego, California

INTRODUCTION

Placing personnel into jobs to maximize expected performance[1] of the organization is a basic problem in large organizations. The solution to this problem requires prediction of the expected performance of each person on each job. These estimates are frequently obtained by developing a separate performance prediction system for each job category.

The predictors in these separate systems consist of information about each person (e.g., age, aptitude scores, interests, experience). After the predictions are made for each person on every possible job, it is desirable to assign each person to a job to maximize expected future performance. This can be accomplished by one of several available computing algorithms (Langley, Kennington, & Shetty, 1974).

If it is necessary to use different sets of prediction weights to make accurate predictions for the various jobs then there is interaction among the people and jobs, and it is important to pay careful attention to the assignment[2] process. However, if it is possible to predict performance accurately using the same set of weights for all jobs, then all possible assignments of personnel to jobs will yield the same overall average performance.

The importance of interaction between people and jobs has been described by Ward (1983). Recognition of the significance of interaction in the predicted payoff array highlights the fact that a constant can be added (or subtracted) from any row or column of the person-job predicted payoff array without changing the particular configuration of assignments of persons to jobs which maximizes the payoff.

---

[1] We make no distinction among productivity, payoff, and performance.

[2] Assignment refers to a general class of personnel actions that includes classification into alternative career fields or job types, assignment to specific job position or location, and other actions such as rotation from one billet to another.

By recognizing that the prediction equations can consist of two types of terms--those that represent the interaction of persons with jobs and those that are additive--we can refer to one set of predictor variables as interactive variables and the other as additive (or noninteractive) variables. Since the noninteractive variable terms can be removed from the operational prediction equations without limiting the assignment process, there is no requirement to have these variables available in calculating predicted payoffs for the optimal assignment of people to jobs. These noninteractive variables are required only to develop the prediction equations in conjunction with the interactive variables. When noninteractive variables increase the amount of interaction (i.e., differential classification potential) of the interactive terms we refer to these noninteractive variables as catalytic variables.[3] Catalytic variables are needed only to develop the weights to be used by the interactive variables, but are not required for making optimal assignments of people to jobs. Therefore, variables can be considered as potential catalytic variables when there is reason to believe that, when they are added to the prediction system in a noninteractive way, they may increase predictive accuracy and increase the person-job interaction and that there is good reason to consider eliminating them from the operational prediction equations. Candidates for catalytic variables are:

1. Variables that have been used operationally but must be eliminated because time is not available to collect the variables. For example, if it is necessary to reduce testing time for the ASVAB, it might be possible to use some subtests as catalytic variables for the others without loss of classification effectiveness. These catalytic subtests would be used in a noninteractive way to determine the weights for the interactive (or operational) subtests. The catalytic subtests would not be required for operational administration to new applicants.

---

[3]The interaction by which we differentiate catalytic variables from interactive variables is between predictor variables and jobs (i.e., of variables represented in a set of regression equations to predict performance in several jobs, those having similar weights for the different jobs are catalytic; those having different weights for the different jobs are interacting). This interaction is contrasted with that occurring in the case of moderator variables where the interaction is between sets of predictor variables (i.e., the weights assigned to one set of predictors are a function of the values for the other set of variables (moderator) (Sanders, 1956). Suppressor variables, on the other hand, are variables which are not themselves significantly correlated with the criterion (job) variables, but which are significantly correlated with other predictor variables which are correlated with the criterion. These variables then "suppress" or control for predictor variance not related to the criterion variable(s) (Horst, 1941). Suppressor variables and catalytic variables are similar in that they both effect a change in the weights assigned other predictor variables when they enter the equation.

2. Variables that have been used experimentally but will not be used operationally. For example, the Vocational Interest Career Examination (VOICE) has been administered to Air Force personnel in conjunction with the Armed Services Vocational Aptitude Battery (ASVAB). Although the VOICE variables are not used operationally, the classification value of the ASVAB might be enhanced by using the VOICE scores as catalytic variables.

3. Some predictor variables may be very expensive. These variables may be collected on a small number of subjects in conjunction with less expensive interactive (operational) variables. The expensive variables can be used as catalytic variables to enhance the operational variables. Therefore, cost of the expensive variables is eliminated.

CATALYTIC VARIABLE CONCEPT

Description of Available Information

Assume that information is available for performance (on the job or at a school) for many individuals on many different jobs and that each person has performed on one and only one job. Also, assume that the same predictor information is available for all persons and that all performance measures are in the same units.

Let

Y_{ij} = the observed performance of person i on job j ($i = 1,\ldots, I_j$ and $j = 1,\ldots, J$).

${}_jX_{ik}$ = the observed value for interactive predictor variable k for person i who has performance Y_{ij} on job j ($k = 1,\ldots, K$).

${}_jC_{i\ell}$ = the observed value for potential Catalytic[4] predictor variable ℓ for person i who has performance Y_{ij} on job j ($\ell = 1,\ldots, L$).

U = a vector of 1s with dimension $I_1 + I_2 + \ldots + I_J = N$, the total number of individuals for whom criterion information has been obtained.

This information is shown in the arrays in Table 1.

---

[4]Catalytic variables will be more formally defined in a later section.

Table 1
The Observed Information

Performance Data Y_{ij}						U_i	Interactive Predictors $_jX_{ik}$					Catalytic Predictors $_jC_{i\ell}$				
1	2	...	j	...	J		1	2	... k	...	K	1	2	... ℓ	...	L
Y_{11}	–	...	–	...	–	1	$_1X_{11}$	$_1X_{12}$	... $_1X_{1k}$	...	$_1X_{1K}$	$_1C_{11}$	$_1C_{12}$	... $_1C_{1\ell}$	...	$_1C_{1L}$
Y_{21}	–	...	–	...	–	1	$_1X_{21}$	$_1X_{22}$	... $_1X_{2k}$	...	$_1X_{2K}$	$_1C_{21}$	$_1C_{12}$	... $_1C_{1\ell}$	...	$_1C_{1L}$
.	.	...	.	...	.	.	.	.	... .	...	.	.	.	... .	...	.
.	.	...	.	...	.	.	.	.	... .	...	.	.	.	... .	...	.
.	.	...	.	...	.	.	.	.	... .	...	.	.	.	... .	...	.
$Y_{i_1 1}$	–	...	–	...	–	1	$_1X_{i_1 1}$	$_1X_{i_1 2}$	... $_1X_{i_1 k}$	...	$_1X_{i_1 K}$	$_1C_{i_1 1}$	$_1C_{i_1 2}$	... $_1C_{i_1 \ell}$	...	$_1C_{i_1 L}$
–	Y_{12}	...	–	...	–	1	$_2X_{11}$	$_2X_{12}$	... $_2X_{1k}$	...	$_2X_{1K}$	$_2C_{11}$	$_2C_{12}$	... $_2C_{1\ell}$	...	$_2C_{1L}$
–	Y_{22}	...	–	...	–	1	$_2X_{21}$	$_2X_{22}$	... $_2X_{2k}$	...	$_2X_{2K}$	$_2C_{21}$	$_2C_{22}$	... $_2C_{2\ell}$	...	$_2C_{2L}$
.	.	...	.	...	.	.	.	.	... .	...	.	.	.	... .	...	.
.	.	...	.	...	.	.	.	.	... .	...	.	.	.	... .	...	.
.	.	...	.	...	.	.	.	.	... .	...	.	.	.	... .	...	.
–	$Y_{i_2 2}$	...	–	...	–	1	$_2X_{i_2 1}$	$_2X_{i_2 2}$	... $_2X_{i_2 k}$	...	$_2X_{i_2 K}$	$_2C_{i_2 1}$	$_2C_{i_2 2}$	... $_2C_{i_2 \ell}$	...	$_2C_{i_2 L}$
–	–	...	Y_{1j}	...	–	1	$_jX_{11}$	$_jX_{12}$	... $_jX_{1k}$	...	$_jX_{1K}$	$_jC_{11}$	$_jC_{12}$	... $_jC_{1\ell}$	...	$_jC_{1L}$
–	–	...	Y_{2j}	...	–	1	$_jX_{21}$	$_jX_{22}$	... $_jX_{2k}$	...	$_jX_{2K}$	$_jC_{21}$	$_jC_{22}$	... $_jC_{2\ell}$	...	$_jC_{2L}$
.	.	...	.	...	.	.	.	.	... .	...	.	.	.	... .	...	.
.	.	...	.	...	.	.	.	.	... .	...	.	.	.	... .	...	.
.	.	...	.	...	.	.	.	.	... .	...	.	.	.	... .	...	.
–	–	...	$Y_{i_j j}$	...	–	1	$_jX_{i_j 1}$	$_jX_{i_j 2}$	... $_jX_{i_j k}$	...	$_jX_{i_j K}$	$_jC_{i_j 1}$	$_jC_{i_j 2}$	... $_jC_{i_j \ell}$	...	$_jC_{i_j L}$
–	–	...	–	...	Y_{1J}	1	$_JX_{11}$	$_JX_{12}$	... $_JX_{1k}$	...	$_JX_{1K}$	$_JC_{11}$	$_JC_{12}$	... $_JC_{1\ell}$	...	$_JC_{1L}$
–	–	...	–	...	Y_{2J}	1	$_JX_{21}$	$_JX_{22}$	... $_JX_{2k}$	...	$_JX_{2K}$	$_JC_{21}$	$_JC_{22}$	... $_JC_{2\ell}$	...	$_JC_{2L}$
.	.	...	.	...	.	.	.	.	... .	...	.	.	.	... .	...	.
.	.	...	.	...	.	.	.	.	... .	...	.	.	.	... .	...	.
.	.	...	.	...	.	.	.	.	... .	...	.	.	.	... .	...	.
–	–	...	–	...	$Y_{i_J J}$	1	$_JX_{i_J 1}$	$_JX_{i_J 2}$	... $_JX_{i_J k}$	...	$_JX_{i_J K}$	$_JC_{i_J 1}$	$_JC_{i_J 2}$	... $_JC_{i_J \ell}$	...	$_JC_{i_J L}$

Notice that in the Y_{ij} array, a dash, –, indicates unknown performance information, since each person performs in one and only one job.

Developing Prediction Equations From Interacting Variables

To determine the least squares regression weights in the usual manner, these data can be used to define the vectors (see Table 2) of N elements ($N = I_1 + I_2 + \ldots + I_J$):

- Y = a vector containing the observed performance Y_{ij}.
- U(j) = a vector with elements equal to 1 if the corresponding element of Y involves job j, 0 otherwise.
- X(jk) = a vector with elements having a value for variable k if the corresponding element of Y is from job j, 0 otherwise.
- E(1) = an error vector.

In this report, symbols in parentheses following a capital letter are used to distinguish vectors (e.g., U(j) (see Table 2) is a vector with elements equal to 1 if the corresponding element of Y is from job j or equal to 0 otherwise; X(jk) is a vector with elements equal to the value for variable k if an element of Y is from job j or equal to 0 otherwise; E(1) is an error vector for Model 1).

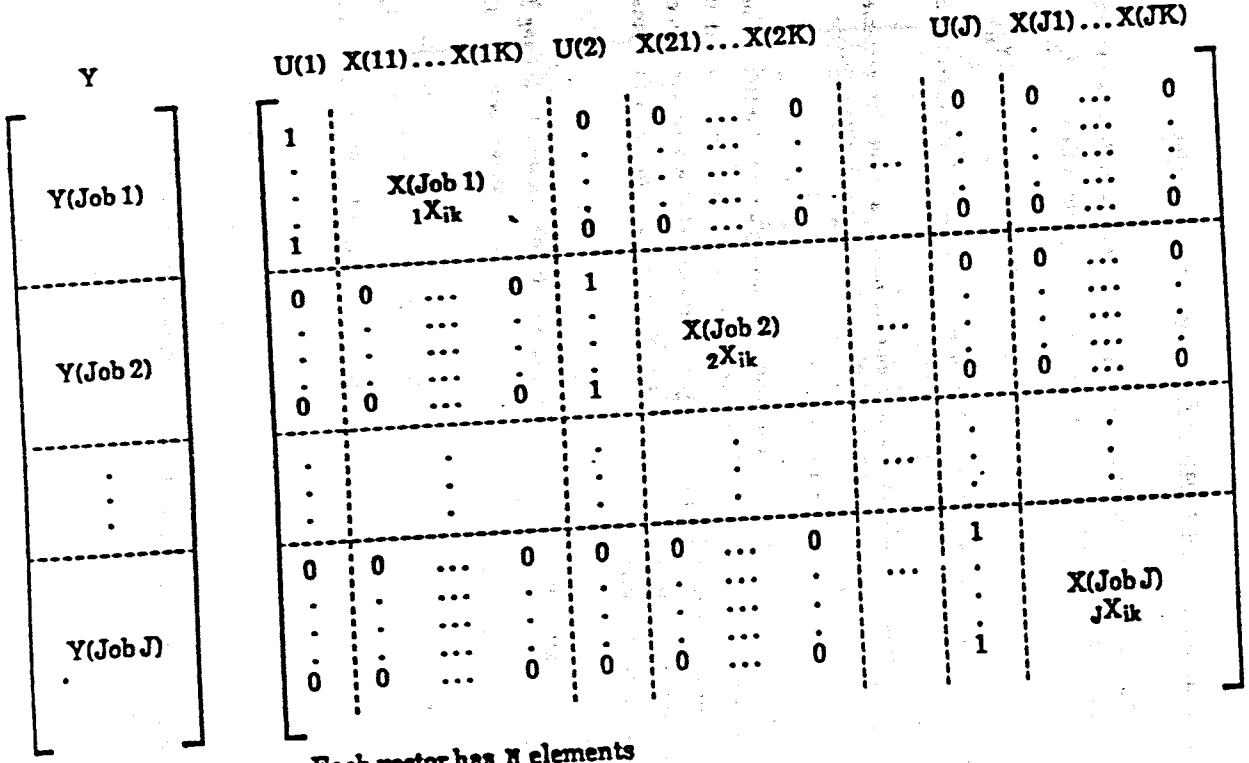
The regression equation coefficients can be determined by solving for the coefficients A_j, B_{jk} for $j=1, \ldots, J$, $k=1, \ldots, K$ in Model 1 shown below. $J(K+1)$ regression coefficients are in the model.

$$
\begin{aligned}
Y = \; & A_1U(1) + B_{11}X(11) + B_{12}X(12) + \ldots + B_{1k}X(1k) + \ldots + B_{1K}X(1K) \\
+ \; & A_2U(2) + B_{21}X(21) + B_{22}X(22) + \ldots + B_{2k}X(2k) + \ldots + B_{2K}X(2K) \\
+ \; & \ldots \\
+ \; & A_jU(j) + B_{j1}X(j1) + B_{j2}X(j2) + \ldots + B_{jk}X(jk) + \ldots + B_{jK}X(jK) \\
+ \; & \ldots \\
+ \; & A_JU(J) + B_{J1}X(J1) + B_{J2}X(J2) + \ldots + B_{Jk}X(Jk) + \ldots + B_{JK}X(JK) + E(1).
\end{aligned}
$$

This single regression model determines a prediction equation for performance on each job from information on the predictor variables (X variables). However, the regression equation for each different job can be computed separately since the vectors associated with each job are orthogonal to the set of vectors associated with each and every other job.

The regression coefficients can be displayed in the array shown in Table 3.

Table 2
Vectors for Determining the Regression Coefficients



Each vector has N elements

Table 3

The Array of Regression Coefficients

A	B			
A_1	B_{11}	B_{12}	...	B_{1K}
A_2	B_{21}	B_{22}	...	B_{2K}
.	.	.	...	.
.	.	.	...	.
.	.	.	...	.
A_J	B_{J1}	B_{J2}	...	B_{JK}

Using the Prediction Equations

After the prediction coefficients have been computed, they can be applied to the predictor information for future groups of personnel to predict future performance for each person on every job. The prediction equations should be applied to a set of people whose data were not used to calculate the regression coefficients. This analysis indicates the degree of confidence that should be placed in future predictors. Since Brogden (1955) has shown that for any assignment of people to jobs, the sum of the multiple regression criterion estimates equals the sum of the actual criterion scores, a further evaluation of the prediction equations can involve comparison of the average performance estimates with that performance from alternative assignments.

Once we have confidence in the prediction equations, the regression coefficients can be applied to a set of data obtained for a total of M subjects (see Table 4).

Let

U = a column vector of 1s of dimension M.

X = a matrix of predictor variables of dimensions M by K.

A = a column vector of regression coefficients of dimension J.

B = a matrix of regression coefficients of dimensions J by K.

A' = the transpose of A.

B' = the transpose of B.

The data set could be new or the same set upon which the prediction equation was developed, in which case $M = N = I_1 + I_2 + \ldots + I_J$.

Table 4

Predicted Performance Array

$$P_{(M\times J)} = \left[\begin{array}{c|c} U & X \\ (M\times 1) & (M\times K) \end{array}\right] \left[\begin{array}{c} A' \\ (1\times J) \\ \hline B' \\ (K\times J) \end{array}\right]$$

$$\begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1J} \\ P_{21} & P_{22} & \cdots & P_{2J} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ P_{M1} & P_{M2} & \cdots & P_{MJ} \end{bmatrix} = \left[\begin{array}{c|ccc} 1 & X_{11} & X_{12} & \cdots & X_{1K} \\ 1 & X_{21} & X_{22} & \cdots & X_{2K} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & X_{M1} & X_{M2} & \cdots & X_{MK} \end{array}\right] \left[\begin{array}{ccccc} A_1 & A_2 & \cdots & A_J \\ \hline B_{11} & B_{21} & \cdots & B_{J1} \\ \cdot & & \cdots & \cdot \\ \cdot & & \cdots & \cdot \\ \cdot & & \cdots & \cdot \\ B_{1K} & B_{2K} & \cdots & B_{JK} \end{array}\right]$$

Table 4 represents the computation of the predicted score matrix P of dimensions M by J. The predicted performance array P can be input into an optimization algorithm to assign persons to jobs to maximize total overall system performance.

Interaction Between Predictor Information and Jobs

It is important to observe the characteristics of the predicted performance array, P. If there is "no-interaction" between the people and jobs, then it makes no difference which persons are assigned to which jobs (Ward, 1983). "No-interaction" conditions between people and jobs in the array, P, means that

$$P_{rt} - P_{ru} = P_{st} - P_{su} = V_{tu} \text{ (a common value) for } r=1,\ldots M-1;\ s=r+1,\ldots,M;\ t=1,\ldots,J-1;\ u=t+1,\ldots,J$$

This can be written as

$$P_{rt} = P_{ru} + V_{tu}$$

and

$$P_{st} = P_{su} + V_{tu}$$

But the conditions for "no-interaction" are equivalent to

$$P_{rt} + P_{su} = P_{ru} + P_{st}$$

This indicates that the sum of the predicted performance values will be the same for all possible assignments of people to jobs.

The conditions for "no-interaction" imply that the regression weights for the corresponding predictors could be identical across all jobs (Ward, 1973, p. 143). It is very important to recognize that even though the weights for the corresponding predictors could be identical across all jobs and have the "no-interaction" conditions in P, it is not necessary that the corresponding weights be identical. For if there is linear dependence among the predictor vectors for a particular job, then there could be an infinite set of weights that would produce the same predicted values for that particular job. It is not possible, in general, to estimate the "amount of interaction" by examining the differences among the corresponding regression coefficients across all jobs.

On the other hand, if the "no-interaction" conditions are not true, it is said that there is "interaction" between the people and the jobs. If there is a "large amount" of interaction, then it is important to seek more optimal assignments. In the presence of such interaction, random assignments could result in extremely poor overall predicted performance. The amount of interaction can be investigated by imposing the restrictions

of "no-interaction" on the prediction systems and examining the loss of predictive accuracy (error sum of squares) when using a single set of weights for all jobs. Imposing the restrictions for "no-interaction" will be discussed in the following section.

No-Interaction Situation

Assume that the "no-interaction" conditions are true for the predicted scores obtained from Model I. This would be the case if:

$$
\begin{array}{ccccccc}
B_{11} & = & B_{21} & = \cdots = & B_{J1} & = & B_1 \\
B_{12} & = & B_{22} & = \cdots = & B_{J2} & = & B_2 \\
\vdots & & \vdots & \cdots & \vdots & & \vdots \\
B_{1k} & = & B_{2k} & = \cdots = & B_{Jk} & = & B_k \\
\vdots & & \vdots & \cdots & \vdots & & \vdots \\
B_{1K} & = & B_{2K} & = \cdots = & B_{JK} & = & B_K.
\end{array}
$$

Since this is never exactly true for real data, we can obtain some indication of the extent of interaction by imposing these restrictions on Model I and obtain the restricted no-interaction regression model, Model Ir:

$$
\begin{aligned}
Y = \quad & A_1U(1) + B_1X(11) + B_2X(12) + \cdots + B_kX(1k) + \cdots + B_KX(1K) \\
+ \quad & A_2U(2) + B_1X(21) + B_2X(22) + \cdots + B_kX(2k) + \cdots + B_KX(2K) \\
+ \quad & \cdots \\
+ \quad & A_jU(j) + B_1X(j1) + B_2X(j2) + \cdots + B_kX(jk) + \cdots + B_KX(jK) \\
+ \quad & \cdots \\
+ \quad & A_JU(J) + B_1X(J1) + B_2X(J2) + \cdots + B_kX(Jk) + \cdots + B_KX(JK) + E(1r),
\end{aligned}
$$

which can be simplified to

$$
\begin{aligned}
Y = \quad & A_1U(1) + A_2U(2) + \ldots + A_jU(j) + \ldots + A_JU(J) \\
& + B_1(X(11) + X(21) + \ldots + X(j1) + \ldots + X(J1)) \\
& + B_2(X(12) + X(22) + \ldots + X(j2) + \ldots + X(J2)) \\
& + \ldots \\
& + B_k(X(1k) + X(2k) + \ldots + X(jk) + \ldots + X(Jk)) \\
& + \ldots \\
& + B_K(X(1K) + X(2K) + \ldots + X(jK) + \ldots + X(JK)) + E(1r).
\end{aligned}
$$

Letting

$$
\begin{aligned}
X(1) &= X(11) + X(21) + \ldots + X(j1) + \ldots + X(J1) \\
X(2) &= X(12) + X(22) + \ldots + X(j2) + \ldots + X(J2) \\
&\ldots \\
X(k) &= (1k) + X(2k) + \ldots + X(jk) + \ldots + X(Jk) \\
&\ldots \\
X(K) &= X(1K) + X(2K) + \ldots + X(jK) + \ldots + X(JK),
\end{aligned}
$$

gives Model 1r

$$
\begin{aligned}
Y = \quad & A_1U(1) + A_2U(2) + \ldots + A_jU(j) + \ldots + A_JU(J) \\
& + B_1X(1) + B_2X(2) + \ldots + B_kX(k) + \ldots + B_KX(K) + E(1r).
\end{aligned}
$$

If the sum of squares of the elements of restricted model error vector E(1r) is significantly larger than the sum of squares of the error E(1), then interaction exists. However, if no interaction (or a "small amount" of interaction) exists, it makes no (or little) difference which people are assigned to which jobs. To observe this, consider assigning any two persons, r and s, to any two jobs, say t and u. Under the assumptions that the prediction weights are identical, the predicted scores will be:

$$
\begin{aligned}
P_{rt} &= A_t + B_1X_{r1} + B_2X_{r2} + \ldots + B_KX_{rK} \\
P_{su} &= A_u + B_1X_{s1} + B_2X_{s2} + \ldots + B_KX_{sK} \\
P_{ru} &= A_u + B_1X_{r1} + B_2X_{r2} + \ldots + B_KX_{rK} \\
P_{st} &= A_t + B_1X_{s1} + B_2X_{s2} + \ldots + B_KX_{sK}.
\end{aligned}
$$

The total predicted performance of assigning person r to job t and person s to job u is the same as assigning person r to job u and person s to job t:

$$P_{rt} + P_{su} = P_{ru} + P_{st}$$

$$A_t + A_u + B_1 (X_{r1} + X_{s1}) + \ldots + B_K (X_{rK} + X_{sK}) =$$
$$A_t + A_u + B_1 (X_{r1} + X_{s1}) + \ldots + B_K (X_{rK} + X_{sK}).$$

It is necessary to have a large amount of interaction between people and jobs in order for alternative assignments to improve the total predicted performance. It is desirable to have a prediction system that provides accurate performance prediction and maintains a large amount of interaction between people and jobs. This observation leads to consideration of catalytic variables.

Introducing Catalytic Variables

In some situations it is possible to add new predictor information that will increase the accuracy of performance prediction, and also increase the amount of interaction between people and jobs. However, requiring additional predictor information can be expensive, difficult, or in some cases quite controversial. Therefore, it would be desirable to add additional predictor information on a small sample that would be required only for development of the prediction equations. But the new information would not be required for future operational assignment of people to jobs. Predictor variables that increase interaction but are not required for future operational use are referred to as catalytic variables.

Catalytic variables were identified in the introduction without definition. They are shown in Table 1 and are designated (as described above) by:

$_jC_{i\ell}$ = the observed value for a potential catalytic predictor variable ℓ for person i who has performance Y_{ij} on job j ($\ell = 1, \ldots, L$).

We will augment Model 1 with the catalytic variables, but require that the coefficients associated with these variables be identical across all jobs. New vectors can be defined:

$C(j\ell)$ = a vector with elements having a value for catalytic variable ℓ if the person performed in job j; 0 otherwise.

Then, Model 2 can be written as:

$$
\begin{aligned}
Y = \quad & A_1U(1) + B_{11}X(11) + B_{12}X(12) + \ldots + B_{1k}X(1k) + \ldots + B_{1K}X(1K) \\
+ \quad & A_2U(2) + B_{21}X(21) + B_{22}X(22) + \ldots + B_{2k}X(2k) + \ldots + B_{2K}X(2K) \\
+ \quad & \ldots \\
+ \quad & A_jU(j) + B_{j1}X(j1) + B_{j2}X(j2) + \ldots + B_{jk}X(jk) + \ldots + B_{jK}X(jK) \\
+ \quad & \ldots \\
+ \quad & A_JU(J) + B_{J1}X(J1) + B_{J2}X(J2) + \ldots + B_{Jk}X(Jk) + \ldots + B_{JK}X(JK) \\
+ \quad & W_1C(11) + W_2C(12) + \ldots + W_\ell C(1\ell) + \ldots + W_LC(1L) \\
+ \quad & W_1C(21) + W_2C(22) + \ldots + W_\ell C(2\ell) + \ldots + W_LC(2L) \\
+ \quad & \ldots \\
+ \quad & W_1C(j1) + W_2C(j2) + \ldots + W_\ell C(j\ell) + \ldots + W_LC(jL) \\
+ \quad & \ldots \\
+ \quad & W_1C(J1) + W_2C(J2) + \ldots + W_\ell C(J\ell) + \ldots + W_LC(JL) + E(2),
\end{aligned}
$$

Where W_ℓ is the coefficient associated with catalytic predictor ℓ for all jobs $j = 1, \ldots, J$. Also, Model 2 can be rewritten as:

$$
\begin{aligned}
Y = \quad & A_1U(1) + B_{11}X(11) + B_{12}X(12) + \ldots + B_{1k}X(1k) + \ldots + B_{1K}X(1K) \\
+ \quad & A_2U(2) + B_{21}X(21) + B_{22}X(22) + \ldots + B_{2k}X(2k) + \ldots + B_{2K}X(2K) \\
+ \quad & \ldots \\
+ \quad & A_jU(j) + B_{j1}X(j1) + B_{j2}X(j2) + \ldots + B_{jk}X(jk) + \ldots + B_{jK}X(jK) \\
+ \quad & \ldots \\
+ \quad & A_JU(J) + B_{J1}X(J1) + B_{J2}X(J2) + \ldots + B_{Jk}X(Jk) + \ldots + B_{JK}X(JK) \\
+ \quad & W_1(C(11) + C(21) + \ldots + C(j1) + \ldots + C(J1)) \\
+ \quad & W_2(C(12) + C(22) + \ldots + C(j2) + \ldots + C(J2)) \\
+ \quad & \ldots \\
+ \quad & W_\ell(C(1\ell) + C(2\ell) + \ldots + C(j\ell) + \ldots + C(J\ell)) \\
+ \quad & \ldots \\
+ \quad & W_L(C(1L) + C(2L) + \ldots + C(jL) + \ldots + C(JL)) + E(2).
\end{aligned}
$$

Define the new vectors

$$C(1) = C(11) + C(21) + \ldots + C(j2) + \ldots + C(J2)$$

$$\vdots$$

$$C(\ell) = C(1\ell) + C(2\ell) + \ldots + C(j\ell) + \ldots + C(J\ell)$$

$$\vdots$$

$$C(L) = C(1L) + C(2L) + \ldots + C(jL) + \ldots + C(JL).$$

Then Model 2 can be written as:

$$\begin{aligned} Y = {} & A_1U(1) + B_{11}X(11) + B_{12}X(12) + \ldots + B_{1k}X(1k) + \ldots + B_{1K}X(1K) \\ & + A_2U(2) + B_{21}X(21) + B_{22}X(22) + \ldots + B_{2k}X(2k) + \ldots + B_{2K}X(2K) \\ & + \ldots \\ & + A_jU(j) + B_{j1}X(j1) + B_{j2}X(j2) + \ldots + B_{jk}X(jk) + \ldots + B_{jK}X(jK) \\ & + \ldots \\ & + A_JU(J) + B_{J1}X(J1) + B_{J2}X(J2) + \ldots + B_{Jk}X(Jk) + \ldots + B_{JK}X(JK) \\ & + W_1C(1) + W_2C(2) + \ldots + W_\ell C(\ell) + \ldots + W_LC(L) + E(2). \end{aligned}$$

There are now J(K+1)+L regression coefficients to be computed. Notice that the new vectors C(1), C(2),. . .,C(L) are <u>not</u> orthogonal to any of the vectors used in Model 1. Therefore, the computational procedure for Model 2 is more complex than for Model 1.

The regression coefficients can be applied from Model 2 either to the data set from which the coefficients were derived or a new data set by augmenting the matrices X, A, and B with the two matrices

C = a matrix of potential catalytic predictor variables designated as $_jC_{i\ell}$ in Table 1.

W = a matrix of regression coefficients of dimension J by L with elements defined as shown below in Table 5 (i.e., the rows are identical).

W' = the transpose of W.

Then, a matrix of predicted values, Q, of dimension M by J, can be obtained as shown in Table 6.

Table 5

Regression Coefficients for Catalytic Variables

$$W = \begin{bmatrix} W_1 & W_2 & \cdots & W_L \\ W_1 & W_2 & \cdots & W_L \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ W_1 & W_2 & \cdots & W_L \end{bmatrix}$$

Observing Predicted Scores From Model 2

Consider again assigning any two persons r and s to jobs t and u. Then, the four predicted scores from Model 2 are:

$$\begin{aligned} Q_{rt} &= A_t + B_{t1}X_{r1} + B_{t2}X_{r2} + \cdots + B_{tK}X_{rK} \\ &\quad + W_1C_{r1} + W_2C_{r2} + \cdots + W_LC_{rL}. \\ Q_{su} &= A_u + B_{u1}X_{s1} + B_{u2}X_{s2} + \cdots + B_{uK}X_{sK} \\ &\quad + W_1C_{s1} + W_2C_{s2} + \cdots + W_LC_{sL}. \\ Q_{ru} &= A_u + B_{u1}X_{r1} + B_{u2}X_{r2} + \cdots + B_{uK}X_{rK} \\ &\quad + W_1C_{r1} + W_2C_{r2} + \cdots + W_LC_{rL}. \\ Q_{st} &= A_t + B_{t1}X_{s1} + B_{t2}X_{s2} + \cdots + B_{tK}X_{sK} \\ &\quad + W_1C_{s1} + W_2C_{s2} + \cdots + W_LC_{sL}. \end{aligned}$$

It can be observed that the difference between the two sums resulting from two different assignments, person r to job t and s to job u, and a second assignment of, person r to job u and s to job t, is given by:

$$(Q_{rt} + Q_{su}) - (Q_{ru} + Q_{st}) =$$

$$(B_{t1}X_{r1} + B_{t2}X_{r2} + \cdots + B_{tK}X_{rK} + B_{u1}X_{s1} + B_{u2}X_{s2} + \cdots + B_{uK}X_{sK})$$
$$-(B_{u1}X_{r1} + B_{u2}X_{r2} + \cdots + B_{uK}X_{rK} + B_{t1}X_{s1} + B_{t2}X_{s2} + \cdots + B_{tK}X_{sK}).$$

Table 6

Predicted Performance Array With Catalytic Variables

$$\underset{(M\times J)}{Q} = \left[\begin{array}{c|c|c} \underset{(M\times 1)}{U} & \underset{(M\times K)}{X} & \underset{(M\times L)}{C} \end{array}\right] \begin{bmatrix} \underset{(1\times J)}{A'} \\ \hline \underset{(K\times J)}{B'} \\ \hline \underset{(L\times J)}{W'} \end{bmatrix}$$

$$\begin{bmatrix} Q_{11} & Q_{12} & \cdots & Q_{1J} \\ Q_{21} & Q_{22} & \cdots & Q_{2J} \\ \vdots & \vdots & \cdots & \vdots \\ Q_{M1} & Q_{M2} & \cdots & Q_{MJ} \end{bmatrix} = \left[\begin{array}{c|ccc|ccc} 1 & X_{11} & \cdots & X_{1K} & C_{11} & \cdots & C_{1L} \\ 1 & X_{21} & \cdots & X_{2K} & C_{21} & \cdots & C_{2L} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{M1} & \cdots & X_{MK} & C_{M1} & \cdots & C_{ML} \end{array}\right] \left[\begin{array}{ccc} A_1 & \cdots & A_J \\ \hline B_{11} & \cdots & B_{J1} \\ \vdots & \cdots & \vdots \\ B_{1K} & \cdots & B_{JK} \\ \hline W_1 & \cdots & W_1 \\ W_2 & \cdots & W_2 \\ \vdots & \cdots & \vdots \\ W_L & \cdots & W_L \end{array}\right]$$

The difference between these two payoff scores is determined only by the Bs and the Xs and there is no need to use the As, Ws, and Cs. The estimates of Bs in Model 2 were made using the information from the catalytic variables, Cs. Therefore, it is not necessary to know the values of Cs for making optimum assignments of future groups of people to jobs.

The addition of the new predictors (Cs) will make the interaction between people and jobs in the new array, Q, larger than the person-job interaction in the original array, P. Greater person-job interaction will allow for greater differential assignment potential. A hypothetical example is presented in the next section to illustrate the effect of a catalytic variable.

A Hypothetical Illustration of a Catalytic Variable

Assume that there are four jobs (J = 4), one interactive predictor variable (K = 1), and one catalytic predictor variable (L = 1). The data analysis might produce the following results for Model 1:

$$Y = A_1 U(1) + B_{11} X(11) + A_2 U(2) + B_{21} X(21) + A_3 U(3) + B_{31} X(31) + A_4 U(4) + B_{41} X(41) + E(1).$$

With numerical values for the As and Bs inserted, Model 1 becomes:

$$Y = 6 U(1) + .4 X(11) + 5 U(2) + .2 X(21) + 3 U(3) + .3 X(31) + 1 U(4) + .1 X(41) + E(1).$$

This regression model can be represented graphically as shown in Figure 1.

Adding the catalytic predictor variable C(1) to the prediction system might result in Model 2:

$$Y = A_1 U(1) + B_{11} X(11) + A_2 U(2) + B_{21} X(21) + A_3 U(3) + B_{31} X(31) + A_4 U(4) + B_{41} X(41) + W_1 C(1) + E(2)$$

With numerical values for the As, Bs, and W_1 inserted, Model 2 becomes:

$$\begin{aligned} Y = {} & 0\,U(1) + 1.1\,X\,(11) \\ & + 2\,U(2) + .8\,X\,(21) \\ & + 4\,U(3) + .1\,X\,(31) \\ & + 5\,U(4) + .4\,X\,(41) + 3\,C(1) + E(2). \end{aligned}$$

The regression model can be represented graphically, as shown in Figure 2, when the value of C(1) = 0. All other graphical representations would differ from Figure 2 by the amount 3 C(1).

Now, consider the assignment of one person with an interactive predictor value of 2 and a second person with an interactive predictor value of 8 to jobs 1 and 4. (Any other combination of persons and jobs could have been considered.)

Using Model 1 gives the predicted values:

	Job 1	Job 4
Person with X = 2	$P_{21} = 6(1) + .4(2)$	$P_{24} = 1(1) + .1(2)$
Person with X = 8	$P_{81} = 6(1) + .4(8)$	$P_{84} = 1(1) + .1(8)$

Then, compare the predicted payoff sum obtained from assigning the person with X = 2 to job 1 and the person with X = 8 to job 4 with the predicted payoff sum obtained from assigning the person with X = 8 to job 1 and the person with X = 2 to job 4. Taking the difference gives:

$$\begin{aligned} (P_{21} + P_{84}) - (P_{81} + P_{24}) &= (6(1) + .4(2) + 1(1) + .1(8)) - (6(1) + .4(8) + 1(1) + .1(2)) \\ &= (.4(2) + .1(8)) - (.4(8) + .1(2)) \\ &= .4(2-8) - .1(2-8) \\ &= (.4-.1)(2-8) = -1.8. \end{aligned}$$

Observe that the difference betwen the two sums (-1.8) is determined only by the product of the difference between the B's (.4 and .1) and the difference between the Xs (2 and 8).

Making the same comparison using Model 2 gives the predicted values:

	Job 1	Job 4
Person with X = 2	$P_{21} = 0(1) + 1.1(2) + 3\,C(1,2)$	$P_{24} = 5(1) + .4(2) + 3\,C(1,2)$
Person with X = 8	$P_{81} = 0(1) + 1.1(8) + 3\,C(1,8)$	$P_{84} = 5(1) + .4(8) + 3\,C(1,8)$

Figure 1. Prediction equations using Model 1.
(Without Catalytic Variable)

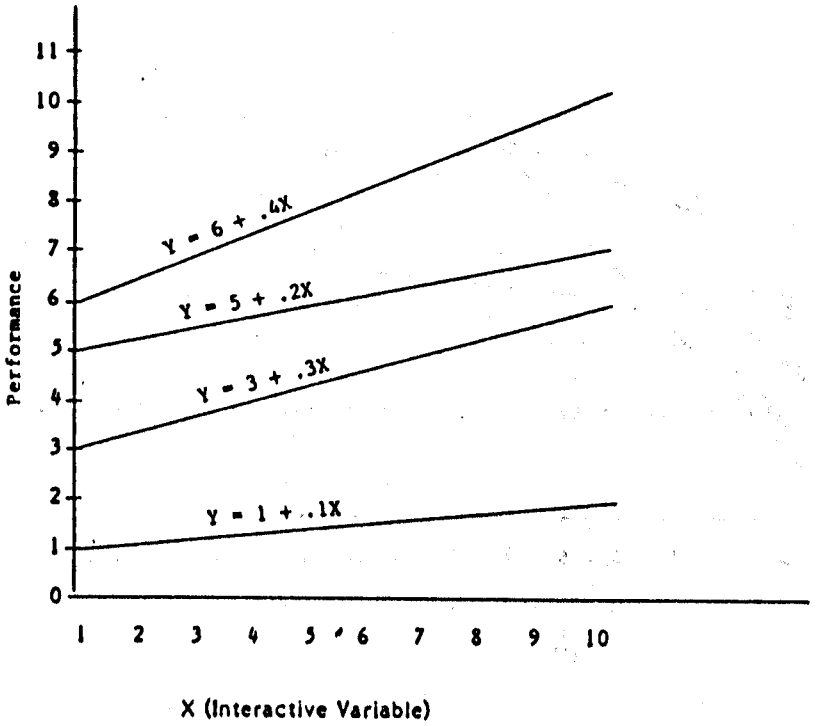
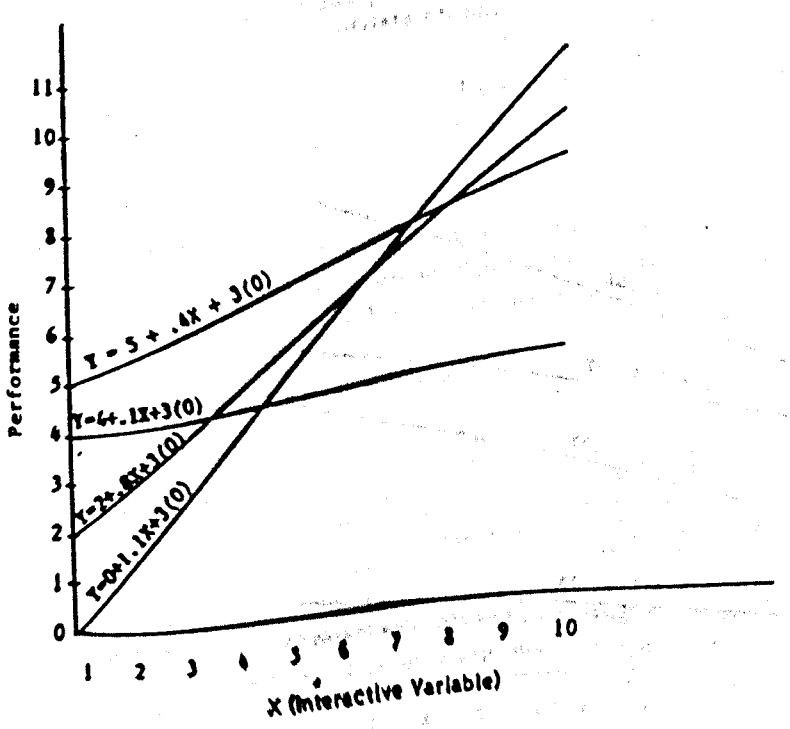


Figure 2. Prediction equations using Model 2. (With Catalytic Variable)



Note - the graph is for C(1) = 0

Then, comparison of the two sums gives the difference

$$
\begin{aligned}
(P_{21} + P_{84}) - (P_{81} + P_{24}) &= (0(1) + 1.1(2) + 3\,C(1,2) + 5(1) + .4(8) + 3\,C(1,8)) \\
&\quad - (0(1) + 1.1(8) + 3\,C(1,8) + 5(1) + .4(2) + 3\,C(1,2)) \\
&= (1.1(2) + .4(8)) - (1.1(1,8) + .4(2)) \\
&= 1.1(2-8) - .4\,(2-8) \\
&= (1.1-.4)\,(2-8) = -4.2.
\end{aligned}
$$

Again the difference between the two sums (-4.2) is determined only by the product of the difference between the Bs (1.1 and .4) and the difference between the Xs (2 and 8). The values of the catalytic weight $W_1 = 3$ and the catalytic values C(1) are not needed for the comparison. The difference (-4.2) using Model 2 is larger absolute value than the difference (-1.8) using Model 1. Comparison of other differences would indicate a tendency for Model 2 differences to be larger than the corresponding differences of Model 1. This would be true because the amount of interaction exhibited in Model 2 is greater than the amount of interaction in Model 1. The comparison of interactions in Model 2 (with) and Model 1 (without) potential catalytic predictors is discussed later.

In this hypothetical illustration, the introduction of the catalytic variable has increased the amount of person-job interaction (and possibly significantly increased predictive accuracy). But having performed its catalytic function, the catalytic variable and its regression weight are no longer required to make optimal assignments that maximize the sum of the predicted performance values.

No-Interaction Situation Using Catalytic Predictors

We can assume no-interaction (i.e., the regression coefficient for each predictor variable is the same for all jobs) and write the same restrictions as before:

$$
\begin{array}{ccccccc}
B_{11} & = & B_{21} & = \cdots = & B_{J1} & = & B_1 \\
B_{12} & = & B_{22} & = \cdots = & B_{J2} & = & B_2 \\
\vdots & & \vdots & \cdots & \vdots & & \vdots \\
B_{1k} & = & B_{2k} & = \cdots = & B_{Jk} & = & B_k \\
\vdots & & \vdots & \cdots & \vdots & & \vdots \\
B_{1K} & = & B_{2K} & = \cdots = & B_{JK} & = & B_K
\end{array}
$$

However, imposing these restrictions on Model 2, we obtain Model 2r:

$$
\begin{aligned}
Y = \; & A_1U(1) + B_1X(11) + B_2X(12) + \ldots + B_kX(1k) + \ldots + B_KX(1K) \\
& + A_2U(2) + B_1X(21) + B_2X(22) + \ldots + B_kX(2k) + \ldots + B_KX(2K) \\
& + \ldots \\
& + A_jU(j) + B_1X(j1) + B_2X(j2) + \ldots + B_kX(jk) + \ldots + B_KX(jK) \\
& + \ldots \\
& + A_JU(J) + B_1X(J1) + B_2X(J2) + \ldots + B_kX(Jk) + \ldots + B_KX(JK) \\
& + W_1C(1) + W_2C(2) + \ldots + W_\ell\ C(\ell) + \ldots + W_LC(L) + E(2r)
\end{aligned}
$$

Simplifying as before we obtain Model 2r:

$$
\begin{aligned}
Y = \; & A_1U(1) + A_2U(2) + \ldots + A_jU(j) + \ldots + A_JU(J) \\
& + B_1X(1) + B_2X(2) + \ldots + B_kX(k) + \ldots + B_KX(K) \\
& + W_1C(1) + W_2C(2) + \ldots + W_\ell\ C(\ell) + \ldots + W_LC(L) + E(2r).
\end{aligned}
$$

If the sum of squares of the elements of restricted model error vector E(2r) is significantly larger than the sum of squares of the error vector E(2), interaction exists. If more interaction exists in Model 2 (when compared to Model 2r) than exists in Model 1 (when compared to Model 1r), the "potential" catalytic predictors may be truly called catalytic.

Comparing Models With and Without Catalytic Predictors

The catalytic effect of predictors that have been added noninteractively to a prediction system can be investigated by comparing the error sum of squares from the four models (1, 1r, 2, and 2r). Alternately, the squared multiple correlations, R_1^2, R_{1r}^2, R_2^2, R_{2r}^2, from the four models can be compared. In each of these models we have:

SSE_1 (sum of squares of error for Model 1) = $N\hat{\sigma}_y^2(1-R_1^2)$,

SSE_{1r} (sum of squares of error for Model 1r) = $N\hat{\sigma}_y^2(1-R_{1r}^2)$,

SSE_2 (sum of squares of error for Model 2) = $N\hat{\sigma}_y^2(1-R_2^2)$, and

SSE_{2r} (sum of squares of error for Model 2r) = $N\hat{\sigma}_y^2(1-R_{2r}^2)$.

Then, computing the differences

$$D_1 = SSE_{1r} - SSE_1 = N\theta_y^2(R_1^2 - R_{1r}^2)$$

$$D_2 = SSE_{2r} - SSE_2 = N\theta_y^2(R_2^2 - R_{2r}^2)$$

provides a basis for examining the catalytic effect of additional predictors. D_1 is the sum of squares associated with interaction without potential catalytic predictors and D_2 is the sum of squares associated with interaction in the presence of potential catalytic predictors.

It is necessary to devise ways to decide if the additional predictor variables have a catalytic effect for differential classification of people to jobs. Observe that D2 is larger than D1 only when $(R_2^2 - R_1^2)$ is greater than $(R_{2r}^2 - R_{1r}^2)$. This means that when the potential catalytic variables are added to the interactive form of the operational variables, they must increase the accuracy of prediction by a larger amount than when they are added to the noninteractive form of the operational variables. Therefore, even if $(R_2^2 - R_1^2)$ is significantly large (i.e., absolute prediction is improved with the addition of the catalytic variables), there could be a decrease in person-job interaction when the potential catalytic variables are added (See Horst (1954, 1955) for discussion of differential vs absolute prediction). In this case, there would be less reason to consider using the additional variables in the catalytic form. On the other hand, if D2 is larger than D1 (and $(R_2^2 - R_1^2)$ is greater than $(R_{2r}^2 - R_{1r}^2)$) we can say that there is an increase in the amount of interaction with the inclusion of the catalytic variables. In such a case we would want to use additional variables in catalytic form.

It is possible to introduce consideration of a super prediction model (Model S) and its squared multiple correlation, R_S^2. This model allows for the investigation of the increase in predictive accuracy and interaction when the potential catalytic variables (Cs) are allowed to have different weights across all jobs (i.e., to join the Xs). Other comparisons among the squared multiple correlations (R_1^2, R_{1r}^2, R_2^2, R_{2r}^2, R_S^2) might be helpful in making decisions about the proper role of the potential catalytic variables. For example:

- if D_2 is much larger then D_1 (indicating increased interaction), and
- if R_2^2 is much larger than R_1^2 (indicating an increase in predictive accuracy), and
- if R_S^2, is insignificantly larger than R_2^2, then we might conclude that the variables would perform very well using only their additive, catalytic form (Model 2).

Further study and experience is needed to develop descriptive, statistical, and practical methods of decision-making about catalytic effects.

APPLICATION OF THE CATALYTIC VARIABLE CONCEPT

The procedure for introducing catalytic predictor variables will be illustrated with data from the military. The first example involves four jobs, one interactive variable (aptitude test) and four potential catalytic variables.

Description of the Information from Example 1

Y_{ij} = Performance measure of individual i on job j:

There are 500 individuals from each job providing a total of 2000 individuals.

$_jX_{ik}$ = the observed interactive predictor (aptitude test score) for individual i who has performance Y_{ij} on job j. (With one interactive variable, k=1.)

$_jC_{il}$ = the observed value for potential catalytic predictor variable ℓ for person i who has performance Y_{ij} on job j (ℓ = 1, 2, 3, 4).

(In the example, each catalytic variable is a mutually exclusive, categorical, binary-coded predictor variable.)

U = a vector of 1s with dimension 2000.

The example data would appear as displayed in Table 7.

Developing Prediction Equations from the Interacting Variable

For the example, the least squares regression weights can be determined in the usual manner by defining the predictor vectors:

Y = a vector containing the observed performance Y_{ij} with N = 2000 elements.

U(j) = 1 if an element of Y is from job j; or 0 otherwise, j = 1, 2, 3, 4.

X(jk) = an ability test value if an element of Y is from job j; or 0 otherwise.

E(1) = an error vector.

Table 7

Observed Information for Example 1

	Performance Data Y_{ij}				U_i	Interactive Predictor $_jX_{1k}$	Catalytic Predictors $_jC_{il}$			
	1	2	3	4			1	2	3	4
	55	-	-	-	1	32	1	0	0	0
	63	-	-	-	1	65	0	0	0	1
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
(I_1 = 500)	82	-	-	-	1	52	1	0	0	0
	-	46	-	-	1	49	0	0	1	0
	-	69	-	-	1	66	0	1	0	0
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
(I_2 = 500)	-	72	-	-	1	38	0	0	0	1
	-	-	62	-	1	53	0	1	0	0
	-	-	93	-	1	59	0	1	0	0
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
(I_3 = 500)	-	-	87	-	1	62	0	0	0	0
	-	-	-	43	1	54	0	0	0	1
	-	-	-	76	1	47	1	0	0	0
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
(I_4 = 500)	-	-	-	82	1	59	0	0	0	1
(N = 2000)										

Notice that the - in the Y_{ij} array indicates unknown performance information, since each person performs in one and only one job. However, the 0 values for the mutually exclusive categorial variables represent nonmembership in the particular category.

Then the regression coefficients can be determined by solving for the regression coefficients A_1, A_2, A_3, A_4, B_{11}, B_{21}, B_{31}, B_{41} in regression Model 1. Observe that K = 1 in this example, since there is only one interactive predictor variable. Model 1 (for example) is:

$$Y = A_1 U(1) + B_{11} X(11) + A_2 U(2) + B_{21} X(21) + A_3 U(3) + B_{31} X(31) + A_4 U(4) + B_{41} X(41) + E(1).$$

As indicated previously, this single regression model determines a prediction equation for performance on each of the four jobs. However, the regression equation for each job can be computed separately since the vectors associated with each job are orthogonal to the set of vectors associated with the other three jobs. The vectors are illustrated in Table 8.

Table 8

Vectors for Determining the Regression Coefficients for Example 1

Y	U(1)	X(11)	U(2)	X(21)	U(3)	X(31)	U(4)	X(41)
55	1	32	0	0	0	0	0	
63	1	65	0	0	0	0	0	
.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	
82	1	52	0	0	0	0	0	
---	---	---	---	---	---	---	---	---
46	0	0	1	49	0	0	0	
69	0	0	1	66	0	0	0	
.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	
72	0	0	1	38	0	0	0	
---	---	---	---	---	---	---	---	---
62	0	0	0	0	1	53	0	
93	0	0	0	0	1	99	0	
.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	
87	0	0	0	0	1	62	0	
---	---	---	---	---	---	---	---	---
43	0	0	0	0	0	0	1	
76	0	0	0	0	0	0	1	
.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	
82	0	0	0	0	0	0	1	

The regression coefficients can be displayed as shown in Table 9.

Using the Prediction Equations for Example 1

The prediction equations can be used to determine the predicted performance of each of M persons on each of the four jobs. The predicted performance matrix P of dimension M by 4 is computed by the matrix multiplication as shown in Table 10.

The predicted performance array P can be put into an optimization algorithm to assign persons to jobs to maximize total system performance. In the example shown, there are only four jobs represented and M people. Usually, there are job quotas for each job such that the sum of the job quotas is equal or very nearly equal to the total number of people to be assigned (M in this case).

Interaction Between Predictor Information (ability test measure) and Jobs

As mentioned above, if there is no interaction between persons and jobs, we would have:

$$B_{11} = B_{21} = B_{31} = B_{41} = B_1.$$

Since this is never exactly true for any real data, some indication of the extent of interaction can be obtained by imposing the restrictions indicated above and solving the restricted no-interaction regression model, Model 1r:

$$\begin{aligned} Y = {} & A_1U(1) + B_1X(11) \\ & + A_2U(2) + B_1X(21) \\ & + A_3U(3) + B_1X(31) \\ & + A_4U(4) + B_1X(41) + E(1r), \end{aligned}$$

which can be simplified to

$$\begin{aligned} Y = {} & A_1U(1) + A_2U(2) + A_3U(3) + A_4U(4) \\ & + B_1(X(11) + X(21) + X(31) + X(41)) + E(1r). \end{aligned}$$

Table 9

The Array of Regression Coefficients for Example 1

$$\begin{bmatrix} A & B \\ A_1 & B_{11} \\ A_2 & B_{21} \\ A_3 & B_{31} \\ A_4 & B_{41} \end{bmatrix}$$

Table 10

Predicted Performance Array for Example 1

$$P_{(M\times 4)} = \left[\begin{array}{c|c} U_{(M\times 1)} & X_{(M\times 1)} \end{array}\right] \begin{bmatrix} A' \\ (1\times 4) \\ \cdots \\ B' \\ (1\times 4) \end{bmatrix}$$

$$\begin{bmatrix} P_1 & P_2 & P_3 & P_4 \\ P_{21} & P_{22} & P_{23} & P_{24} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ P_{M1} & P_{M2} & P_{M3} & P_{M4} \end{bmatrix} = \left[\begin{array}{c|c} 1 & X_{11} \\ 1 & X_{21} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & X_{M1} \end{array}\right] \begin{bmatrix} A_1 & A_2 & A_3 & A_4 \\ \hline B_{11} & B_{21} & B_{31} & B_{41} \end{bmatrix}$$

Letting $X(1) = X(11) + X(21) + X(31) + X(41)$ give Model 1r:

$$Y = A_1U(1) + A_2U(2) + A_3U(3) + A_4U(4) + B_1X(1) + E(1r).$$

If the sum of squares of the elements of restricted model error vector E(1r) is "significantly" (statistically and/or practically) larger than the sum of squares of the error vector E(1) (R^2_{1r} smaller than R^2_1), then interaction exists. If interaction is not indicated, individuals can be assigned (e.g., arbitrarily or randomly) to any job without affecting total predicted performance.

Introducing Catalytic Variables

Catalytic variables were defined earlier as predictor variables that increase interaction between people characteristics (i.e., interacting variables) and jobs, but are not required for future operational use (i.e., do not interact themselves with jobs) to optimally classify people into jobs.

In our example, Model 1 will be augmented with four catalytic predictor variables. However, as indicated earlier, the regression coefficients associated with each of these four catalytic predictor variables must be the same for all four jobs. There should be no interaction between catalytic predictor variables and jobs.

Then, four catalytic predictor vectors can be defined in our example.

- C(1) = a vector for catalytic variable 1, which, in the example, is a binary-coded predictor having a value of 1 if the observation comes from the first mutually exclusive category and 0 otherwise.
- C(2) = a vector for catalytic variable, 2 which, in the example, is a binary-coded predictor having a value of 1 if the observation comes the second mutually exclusive category from and 0 otherwise.
- C(3) = a vector for catalytic variable 3, which is defined similar to C(1) and C(2).
- C(4) = a vector for catalytic variable 4, which is defined similar to C(1), C(2), and C(3).

Then, the final form of regression Model 2 above can be written as:

$$
\begin{aligned}
Y = \; & A_1 U(1) + B_{11} X(11) \\
& + A_2 U(2) + B_{21} X(21) \\
& + A_3 U(3) + B_{31} X(31) \\
& + A_4 U(4) + B_{41} X(41) \\
& + W_1 C(1) + W_2 C(2) + W_3 C(3) + W_4 C(4) + E(2).
\end{aligned}
$$

The predictor vectors C(1), C(2), C(3), and C(4) are generally not orthogonal to the other vectors. Therefore, the computational procedure for Model 2 is more complex than for Model 1. It is important to note that the least squares estimates of the values for A_1, A_2, A_3, A_4, B_{11}, B_{21}, B_{31}, and B_{41} are not generally the same in Models 1 and 2. After solving for the coefficients in Model 2, the predicted performance matrix Q of dimension M by 4 from the matrix multiplication, as shown in Table 11, can be obtained.

No-Interaction Between Predictor Information and Jobs Using Catalytic Predictors

The hypothesis of no-interaction can be investigated as before by assuming in Model 2 that:

$$B_{11} = B_{21} = B_{31} = B_{41} = B_1$$

and imposing these restrictions obtain the restricted model, Model 2r.

$$
\begin{aligned}
Y = \; & A_1 U(1) + B_1 X(11) \\
& + A_2 U(2) + B_1 X(21) \\
& + A_3 U(3) + B_1 X(31) \\
& + A_4 U(4) + B_1 X(41) \\
& + W_1 C(1) + W_2 C(2) + W_3 C(3) + W_4 C(4) + E(2r).
\end{aligned}
$$

As before, if the sum of squares of the elements of restricted model error vector E(2r) is "significantly" (statistically and/or practically) larger than the sum of squares of the error vector E(2) (R^2_{2r} smaller than R^2_2), then it can be concluded that interaction exists. If more interaction exists in Model 2 (when compared to Model 2r) than exists in Model 1 (when compared to Model 1r), the "potential" catalytic predictors can be truly called catalytic.

Table 11

Predicted Performance Array With Catalytic Variables for Example 1

$$\underset{(M\times 4)}{Q} = \left[\begin{array}{c|c|c} \underset{(M\times 1)}{U} & \underset{(M\times 1)}{X} & \underset{(M\times 4)}{C} \end{array}\right] \left[\begin{array}{c} \underset{(1\times 4)}{A'} \\ \hline \underset{(1\times 4)}{B'} \\ \hline \underset{(4\times 4)}{W'} \end{array}\right]$$

$$\begin{bmatrix} Q_{11} & Q_{12} & Q_{13} & Q_{14} \\ Q_{21} & Q_{22} & Q_{23} & Q_{24} \\ \vdots & \vdots & \vdots & \vdots \\ Q_{M1} & Q_{M2} & Q_{M3} & Q_{M4} \end{bmatrix} = \left[\begin{array}{c|c|cccc} 1 & X_{11} & C_{11} & C_{12} & C_{13} & C_{14} \\ 1 & X_{21} & C_{21} & C_{22} & C_{23} & C_{24} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{M1} & C_{M1} & C_{M2} & C_{M3} & C_{M4} \end{array}\right] \left[\begin{array}{cccc} A_1 & A_2 & A_3 & A_4 \\ \hline B_{11} & B_{21} & B_{31} & B_{41} \\ \hline W_1 & W_1 & W_1 & W_1 \\ W_2 & W_2 & W_2 & W_2 \\ W_3 & W_3 & W_3 & W_3 \\ W_4 & W_4 & W_4 & W_4 \end{array}\right]$$

Comparing Interactions With and Without Catalytic Predictors for Examples

As indicated above, the catalytic effect of predictors that have been added noninteractively to a prediction system can be investigated by comparing the error sum of squares from the four models (1, 1r, 2, 2r).

For the example 1 data with N = 2000 we obtained:

$$D_2 - D_1 = N\hat{\sigma}_y^2((R_2^2 - R_{2r}^2) - (R_1^2 - R_{1r}^2))$$

$$D_2 - D_1 = N\hat{\sigma}_y^2((.1930 - .1785) - (.1832 - .1705))$$

$$= N\hat{\sigma}_y^2 (.0018) = 2000\,\hat{\sigma}_y^2 (.0018).$$

The fact that D_2-D_1 is greater than zero indicates that some catalytic effect is due to the four predictors C(1), C(2), C(3), and C(4). Also, the increase in absolute predictive accuracy ($R_2^2 - R_1^2$) is statistically significant. Other information would be required to decide if the catalytic variables are practically useful.

A second random sample of 2000 subjects was chosen and the analysis was repeated. The difference from Sample 2 was:

$$D_2\text{-}D_1 = N\hat{\sigma}_y^2((R_2^2\text{-}R_{2r}^2) - (R_1^2\text{-}R_{1r}^2))$$

$$D_2\text{-}D_1 = N\hat{\sigma}_y^2((.1649\text{-}.1539)\text{-}(.1571\text{-}.1471))$$

$$= N\hat{\sigma}_y^2(.0010) = 2000\,\hat{\sigma}_y^2(.0010).$$

The second sample also indicates statistically significant increase in absolute prediction, and an increase in the amount of interaction (D_2 greater than D_1). This suggests the possibility of using the catalytic variables.

Example of Noncatalytic Effects

Example 2 has been chosen to illustrate "potential" catalytic variables that result in decrease in interaction and, therefore, become noncatalytic variables. This example consists of three jobs, one interactive variable (aptitude test) and 2 potential catalytic variables. There are a total of 7043 people in the example, 2317 subjects from job 1, 1836 subjects from job 2, and 2890 subjects from job 3.

For this example we can compute the difference between the interaction sum of squares without and with potential catalytic variables:

$$D_2 - D_1 = N\hat{\sigma}_y^2((R_2^2\text{-}R_{2r}^2) - (R_1^2\text{-}R_{1r}^2))$$

$$D_2 - D_1 = N\hat{\sigma}_y^3((.3647\text{-}.3623) - (.3385\text{-}.3349)$$

$$D_2 - D_1 = N\hat{\sigma}_y^2(-.0012)$$

$$= 7043\hat{\sigma}_y^2(-.0012).$$

The negative value of D_2-D_1 indicates that there is a decrease in person-job interaction (differential prediction) when the potential catalytic variables are added. However, there is a statistically significant increase in absolute predictive accuracy. It would be doubtful that the addition of the catalytic variables would be of practical value in this case.

Catalytic Effects in Operational Situations

The actual catalytic effect in an operational situation depends on the particular set of people and jobs under consideration. The predicted scores P (without potential catalytic predictors) and the predicted scores Q (with potential catalytic predictors) should be computed for a particular set of people and jobs. The interaction sum of squares for the P matrix (designated D_p) can be compared with the interaction sum of squares for the Q matrix (designated D_q) in the same manner as above. As before, it is suggested that, if D_q is larger than D_p, then, for this particular set of people and jobs, the additional predictor variables have a catalytic effect.

As the interaction between people and jobs increases, it becomes more important to assign the "right person to the right job."

CONCLUSIONS

If there is no interaction between people characteristics and jobs in the prediction of job performance, then it makes no difference in overall system performance which people are assigned to which jobs. To increase interaction (and, therefore, differential assignment potential), it is usually necessary to add new variables to the operational variables in the prediction system. The addition of new variables can be costly, time consuming, and frequently controversial. The approach described herein suggests adding predictor variables in a noninteractive way to the operational (interacting) predictors to increase the possibility of more interaction between people and jobs. If these additional noninteractive variables can increase interaction, they are called catalytic variables. Catalytic variables (which enter the prediction system in an additive way) are not required for use in the assignment of people to jobs to maximize overall system performance.

The statistical and practical significance of the catalytic effects approach should be studied to develop guidelines for making cost-benefit decisions about the use of catalytic variables.

To gain more knowledge about the catalytic process, data already collected for people, jobs, and potential catalytic variables should be studied.

Data sets involving performance measures requiring a wide variety of attributes, and a large number of different jobs should be used to maximize the prospects of finding catalytic predictors.

REFERENCES

Brogden, H. F. (1955). Least squares estimates and optimal classification. Psychometrika, 20(3), 249-252.

Horst, P. (1941). The prediction of personal adjustment. Social Science Research Bulletin, 48.

Horst, P. (1954). A Technique for the development of a differential prediction battery. Psychological Monographs, No. 380.

Horst, P. (1955). A Technique for the development of a multiple absolute prediction battery. Psychological Monographs, No. 390.

Langley, R. W., Kennington, J., & Shetty, C. M. (1974). Efficient computational devices for the capacitated transportation problem. Office of Naval Research, Naval Research Logistics Quarterly, 21(4), 637-647.

Sanders, D. R. (1956). Moderator variables in prediction. Educational and Psychological Measurement, 16, 209-222.

Ward, J. H., Jr. (1983). Strategies for capitalizing on individual differences in military personnel systems. In R. C. Sorenson (Ed.). Human individual differences in military systems (NPRDC Spec. Rep. 83-30). San Diego: Navy Personnel Research and Development Center.

Ward, J. H., Jr., & Jennings, E. (1973). Introduction to linear models. Prentice-Hall Inc., Englewood Cliffs, NJ.

SUMMARY

Organizations have a fundamental problem of placing personnel into jobs to maximize expected performance. Whether or not placing people in specific jobs really makes a difference in overall expected system performance depends on the interaction of people characteristics with jobs. It is desirable to increase the interaction of the people characteristics, as measured by predictor tests, with the jobs.

The purpose of this effort is to suggest a procedure for using one set of performance predictor variables in a simple noninteractive way to enhance the differential classification potential (person-job interaction) of a set of operational predictor variables. The noninteractive variables are required only in determination of the regression coefficients for the operational predictors, but are not required for operational use in future differential classification actions.

Separate equations are developed to predict performance on each job. The equations are determined so that the weights for the operational predictors are allowed (if necessary) to vary across the various jobs. However, one set of predictors (the potential catalytic variables) is required to have the same regression weights across all jobs (noninteractive). If this noninteractive set of predictors can increase the amount of person-job interaction in the new predicted performance values, then the potential for improved assignment has been increased. These noninteractive variables are called catalytic.

Since catalytic variables are used in prediction systems in a noninteractive way, they are not required for future use in the classification system. Therefore, this procedure will allow personnel classification system developers to use a set of catalytic predictors to enhance the differential classification potential of a set of operational (interactive) predictors, but not require these catalytic predictors for future classification. If catalytic variables can be found, savings in time and money might be possible with little loss in classification effectiveness of the operational predictors.

This approach should be applied to prediction situations in which data are already available and it is desirable to enhance the classification effectiveness of a set of operational predictors without requiring the operational use of the catalytic variables.

)LTIPLE LINEAR REGRESSION VIEWPOINTS
)LUME 15, NUMBER 1, SUMMER 1986

Testing Different Model Building Procedures Using Multiple Regression

Jerome D. Thayer
Andrews University

One of the most appealing aspects of multiple regression to beginning multiple regression students is the amazing feat performed by a stepwise regression computer program. The process of selecting the "best" combination of predictors so effortlessly and efficiently creates an overwhelming urge to use this procedure and the computer program that accomplishes it for a multitude of tasks for which it is ill suited. Many textbooks on multiple regression claim that abuse of this technique is common. Draper and Smith (1981) give a mild statement that "the stepwise procedure is easily abused by amateur statisticians (p. 310), while Wilkinson (1984) is much more dramatic:

> Stepwise regression is probably the most abused computerized statistical technique ever devised. If you think you need stepwise regression to solve a particular problem you have, it is almost certain that you do not. Professional statisticians rarely use automated stepwise regression. (p. 196)

Cohen and Cohen (1975) suggest that model building should proceed according to dictates of theory rather than relying on the whims of a computer. But since in the social and behavioral sciences theoretical models are relatively rare (Neter et al., 1983), Cohen and Cohen suggest that the stepwise method is a "sore temptation" to replace theory in these situations (p. 103).

The authors of current multiple regression textbooks suggest the following considerations for selecting a subset of predictors for a regression model:

1. Selection of variables for a regression model should not be a mechanical process (Chatterjee and Price, 1977; Draper and Smith, 1981; Neter et al., 1983; Younger, 1979).

2. No one process will consistently select the "best" model (Berenson et al., 1983; Gunst and Mason, 1980; Kleinbaum and Kupper, 1978; Morrison, 1983; Pedhazur, 1982; Younger, 1979).

3. There is no one "best" model according to any common criterion such as the maximum R^2 (Chatterjee and Price, 1977; Freund and Minton, 1979; Neter et al., 1983).

4. The stepwise method should not be used to build models for explanatory research (Cohen and Cohen, 1975; Pedhazur, 1982).

In addition many authors point out that the stepwise method has limited usefulness when the predictors are highly correlated (Chatterjee and Price, 1977; Kleinbaum and Kupper, 1978; Neter et al., 1983), if a key set of variables work in combination (Younger, 1979), or when suppression exists (Cohen and Cohen, 1975). Chatterjee and Price (1977) suggest that with multicollinearity the backward method is preferred although other authors suggest that the backward method should not be used in this case because of computational inaccuracy that may occur if multicollinearity is severe and a near singular matrix is inverted.

In spite of these suggestions, there are still many research studies reported in the literature in which these guidelines are violated. Results are reported of a model "selected" by the computer, usually using the stepwise method with no indication that this model might not be the "correct" or "best" one. The discussion of the selected model is done in a mechanical fashion with no indication given of a careful critique of the adequacy of the computer-selected model. Explanatory interpretations are frequently made (Pedhazur, 1982) which often take the form of considering variables selected by the computer to be "good" predictors of the dependent variable because they have a "significant relationship" and variables not selected by the computer are considered to be "poor" predictors because they do not have a "significant relationship". A variable that may be one of the best predictors when studied individually and that fits nicely into an existing theory will be considered to be a "poor" predictor simply because it does not occur in the selected model even though its omission may be due to predicting the same variance as

other predictors already in the model that are no better predictors than it is.

There are many other competing procedures that can be used to select variables for a regression model other than the stepwise method. Three major ones mentioned in many regression textbooks are the forward, backward, and best subsets methods. This paper will endeavor to compare the stepwise method with these selection methods to determine the types of models that each would be likely to select and in so doing determine the strengths and weaknesses of each method.

Method

The procedure used was to apply each of the common selection methods to a number of data sets of various types and evaluate the differences between the models chosen. The source for each of the data sets used in the analysis is described below. In Table 1 the number of subjects and number of predictors for each data set is listed.

Data Sets Used

1. GMA1 -- Data Set A1 from Gunst and Mason (1980)
2. GMA3 -- Data Set A3 from Gunst and Mason (1980)
3. GMA6 -- Data Set A6 from Gunst and Mason (1980)
4. GMA8 -- Data Set A8 from Gunst and Mason (1980)
5. GMB1 -- Data Set B1 from Gunst and Mason (1980)
6. GMB2A-GMB2B -- Data Set B2 from Gunst and Mason (1980)
7. TAL -- Project Talent data from Lohnes and Cooley (1968)
8. ENR1-ENR5 -- 1986 freshman enrollment data from Andrews University
9. LONG -- Data from Longley (1967)
10. HALD -- Data from Draper and Smith (1981)
11. SUP -- Data generated from a contrived correlation matrix

Nine of the data sets were selected from textbooks that used the data sets to illustrate interesting and/or unusual applications of regression that would be brought out by the data. All of the variables were not included in some of the sets. Some of the variables in the GMA3 set were not used because there were more variables than subjects. One variable was removed from the GMB1 set due to tolerance problems (its tolerance was below .01, and thus was automatically excluded from the BMDP2R program although it would not have been included in any of the models if tolerance had been ignored). The categorical variables from the TAL set were not used.

The SUP data was generated using a program described in Morris (1975) from a contrived correlation matrix described below that included variables that illustrated suppression. To get a correlation matrix with suppression, three variables were constructed composed of random numbers with the first variable designated as the dependent variable and the other two designated as independent variables. A fourth variable was then constructed which did not have a high correlation with the dependent variable by itself but yielded a high multiple correlation with the dependent variable when combined with the two previously chosen independent variables. The correlation matrix from this data was then used as input to the Morris program which generated a new set of data which gave the same correlation matrix but was "marginally normal." The correlation matrix used was:

	1	2	3	4
1	1.000	.446	.292	.397
2		1.000	-.195	-.088
3			1.000	-.527
4				1.000

An alternate approach that would have given an equivalent matrix would have been to use the method suggested by Lutz (1983).

GMB2 was run twice using a different dependent variable each time. The ENR data was analyzed with 5 different sets of predictors. The variables used for the ENR data sets were selected from 86 variables which in turn were selected from a larger data base that included 499 variables. A principal components factor analysis was conducted using the 86 variables and the variables loading on the 14 factors with the highest eigen values (all above 1.3) were used in the 5 sets of predictors.

ENR1 had 1 predictor from each of the first 7 factors.

ENR2 had 2 predictors from each of the first 7 factors.

ENR3 had 4 predictors from each of the first 7 factors.

ENR4 had 1 predictor from each of the 14 factors.

ENR5 had 2 predictors from each of the 14 factors.

The computer programs used to select the best model from each data set were BMDP2R for the stepwise, forward and backward solutions, and BMDP9R for the best subsets solution. The stepwise and forward methods used an F-to-enter limit of 2.0 and the stepwise method used an F-to-remove limit of 1.99. These limits are in line with recommendations made for proper use of stepwise regression which suggest that the F-to-enter limit selected should be fairly low so as to allow more variables a chance to show their worth in the final model. The backward method used a comparable F-to-remove limit of 2.0. The BMDP9R program selected the model with the lowest C_p value, which is the default value of the program. An ideal C_p value is one that is equal to or lower than the number of parameters in the model (predictors + 1). Dixon and Brown (1979) suggest that this criterion will give models in which the variables in the model have F-to-remove values above 2.0, making this criterion similar to that used in the other three methods. Of course, the specific models selected would differ if other criteria were used, but the overall characteristics of the four selection methods should not change. To evaluate a different criterion, on some comparisons it will be noted what the

results would have been if an F-to-enter/remove level of 4.0 had been used rather than 2.0.

Table 1 reports the characteristics of the subsets selected by the 4 selection methods with the 16 data sets. For the stepwise method the number of predictors selected is reported along with the R^2 for the selected model. For the other methods information is only presented if the model selected was different from the model selected by the stepwise method. Additional information provided for these models includes the number of predictors in that model that were not in the stepwise model and the number of predictors in the stepwise model not included in that model.

Results

On 9 of the 16 data sets, the 4 methods chose different models using the initial criteria of a F-to-enter/remove of 2.0 and the lowest C_p. In comparison with the stepwise method, the forward method chose a different model on 2 data sets, the backward method chose a different model on 5 data sets, and the best subsets method chose a different model on 7 data sets. The backward method and best subsets method differed on 4 data sets. For each of the data sets on which differences were found, the differences will be described in detail.

GMA3 -- The stepwise, backward and best subsets methods selected the same model which had 1 less variable than that selected by the forward method. If F-to-enter/remove limits of 4.0 had been used, the stepwise and backward methods would have removed one additional variable giving a 4 predictor model while the model chosen by the forward method would not have changed, thus having 2 more predictors than the stepwise and backward methods.

GMA6 -- The backward and best subsets methods gave the same model which had an R^2 more than twice as much as that found by the stepwise and forward methods which gave the same model. The R^2 values found were .150 and .347.

The stepwise/forward model had 2 predictors and the backward/best subsets model had 7 predictors. The stepwise/forward methods did not enter a third variable because the highest F-to-enter was 1.96. The worst variable in the 7 variable backward and best subsets model had a F-to-remove of 3.25. If an F-to-enter limit of 4.00 had been used, there would have been no variables included in the stepwise/forward model since the first variable entered had an F-to-enter of 2.50 while the backward method would have removed the seventh variable leaving a 6 variable model with an R^2 of .300. The stepwise method gave much lower R^2 values at F-to-enter limits of both 2.0 and 4.0. The C_p value for the backward/best subsets model was 4.02 for 7 predictors while the stepwise/forward model had a C_p value of 5.54 for 2 predictors, indicating the 7 predictor model chosen by the backward and best subsets methods was a much better model.

GMA8 -- The stepwise, forward, and backward methods produced the same model which was different from that chosen by the best subsets method. The best subsets model had 1 less predictor, the last variable chosen by the stepwise/forward methods and the variable which would have been the next to be deleted by the backward method. The R^2 values for the 2 models were .886 and .877. The C_p values for the 2 models were about identical (1.51 for the stepwise/forward/backward model and 1.50 for the best subsets model). The F-to-remove for the fourth variable included in the larger model was 2.28.

GMB1 --The 4 methods produced 3 models, with the stepwise and forward methods selecting the same model. The R^2 values for the models were .716 for the 5 predictor best subsets model, .727 for the 6 predictor stepwise/forward model, and .739 for the 8 predictor backward model. All of the variables in the best subsets model were included in the stepwise/forward model with the additional variable in the stepwise/forward model having an F-to-enter of 2.02. The backward model used 4 of the 6 predictors in the stepwise/forward model and 4 additional predictors. The C_p values were 3.27 for the

stepwise/forward model and 3.14 for the best subsets model. The backward model was not listed as one of the 10 best 8 predictor models in the BMDP9R best subsets selection even though it had an R^2 of .737 which was higher than 9 of the 8 variable models listed. If the F-to-enter and F-to-remove limits had been 4.0, both the stepwise/forward and backward models would have included 5 variables but only 3 would have been common to both. The 5 variable model R^2 would have been .716 for the stepwise/forward model and .697 for the backward model.

GMB2B -- The model selected by the stepwise and forward methods had only 1 predictor with an R^2 value of .176. No variable was even close to being considered for entry as the F-to-enter value for the best additional second variable was 0.76. The backward and best subsets models were the same with 5 predictors and an R^2 of .509. The worst variable in the 5 predictor model had an F-to-remove value of 6.82. The reason for the discrepancy between the models was that 2 of the variables were only good predictors in combination. In the stepwise solution, one of this pair would have been the second variable added with an F-to-enter of 0.76 and increasing the R^2 from .176 to .193. The third variable added would have been the other member of the pair which would have increased the R^2 to .371. The better predictor of the pair in the second step added only .017 (.193-.176) while together as steps 2 and 3, the pair added .195 (.371-.176). The fourth and fifth predictors increased the R^2 from .371 to .509.

TAL -- All of the methods selected the same model but the order of entry of the variables in the stepwise/forward and backward methods were different. The last variable entered in the stepwise and forward methods was not the same as the variable that would have been removed next in the backward method. If the F-to-enter/remove limit had been 4.0, the models would have been different with the stepwise/forward method model having 4 variables with an R^2 of .388 and the backward model having 6 variables with an R^2 of .396. The additional

2 variables for the backward model were included because these 2 variables would not have been good enough to enter alone in the stepwise/forward methods, but together they were good predictors, making them remain in the backward method.

ENR3 -- The 4 methods produced 3 models, with the stepwise and forward methods selecting the same model. The R^2 values for the models were .520 for the 8 predictor best subsets model, .521 for the 9 predictor stepwise/forward model, and .525 for the 11 predictor backward model. All of the variables in the best subsets model were included in the stepwise model with the additional variable of the stepwise model having an F-to-enter of 2.02. All but one of the variables in the stepwise/forward model were included in the backward model with 3 additional variables added. The 3 models selected were the best, second best, and tied for third best in the best subsets method with C_p values of 5.88, 5.89, and 6.05. The other model with a C_p of 6.05 was the second best 8 predictor model selected by the best subsets method. This model had 1 predictor different from the best model selected. It appears as if the additional 2 or 3 variables of the backward model were not needed to select a good model but other combinations of variables would have given equally good smaller models. If an F-to-enter limit of 4.00 had been used, the stepwise/forward model would have contained 5 predictors with an R^2 of .510 and the backward model would have had 7 predictors with an R^2 of .517 with only 3 of the same predictors as the stepwise/forward model.

ENR5 -- All of the methods produced the same model but the stepwise/forward and backward models had a different order of entry. If the F-to-enter/remove limit had been 4.00, the stepwise/forward model would have had 8 predictors with a R^2 of .338 and the backward model would have had 9 predictors with a R^2 of .343 with 6 variables the same as those in the stepwise/forward model. If the ninth predictor of the backward model had been

removed, the remaining 8 variables would have had the same R^2 as the stepwise/forward model (.338) with 2 variables being different.

LONG -- The stepwise, forward and backward methods chosen by BMDP2R gave the same 3 predictor model with an R^2 of .985 and the best subsets model had 4 predictors with an R^2 of .995. The additional predictor in the best subsets model was not included in the other models due to its high intercorrelation (tolerance=.002) with the first 3 predictors in the model. BMDP9R (best subsets) allows a greater degree of multicollinearity than BMDP2R, so this problem was not encountered with the model chosen by that program. The F-to-remove value of the fourth variable in the best subsets model was 5.95 indicating it deserved to be in the model if the low tolerance could be ignored. The C_p value for the 4 predictor model was 3.24 compared to the 3 predictor value of 21.66. The first variable entered in the stepwise and forward methods was the variable that contributed the most to the high tolerance value for the fourth variable in the model (the correlation between them was .995). If a 3 predictor model had been chosen by all methods ignoring the tolerance problem, the backward and best subset methods would have chosen the same model with a higher R^2 than that chosen by the stepwise/forward method (.993 to .985). The C_p value for the 3 predictor backward/best subsets model would have been 6.24 compared to the stepwise/forward value of 21.66. The backward/best subsets model is better because the second and third variables entered in the stepwise/forward method in combination pair much better with the fourth variable than the first variable entered. The model chosen by the backward and best subsets methods was never evaluated in the stepwise and forward methods.

HALD --The stepwise, backward, and best subsets chose the same 2 predictor model while the forward method selected a 3 predictor model, including a variable that was the first one entered but that later became redundant with the addition of the second and third variables.

SUP -- The stepwise and forward methods did not allow any variables to enter the model. The largest F-to-enter value was 1.99. The backward and best subsets models were the same with 3 predictors and an R^2 of .967. The lowest F-to-remove value of the 3 predictors was 85.16 which if removed would bring the R^2 down to .506. Each variable acting alone did not predict enough to be included but only showed its high predictive power in combination with the other variables.

Conclusions

If models chosen by different selection methods were relatively similar in the number of variables in the model, the variables included, and the amount of variance explained (R^2), and the model was to be used primarily for prediction, not explanatory purposes, it would seem that the suggestion of Draper and Smith (1981) that the stepwise method might be preferred because of its practical nature would seem reasonable. The results of this study suggest, however, that in some cases models that are severely inadequate are selected by the stepwise method and other consistent, but less important differences between the models selected by the different methods also appear.

Forward/stepwise comparison

It would be expected that the forward method would be more similar to the stepwise method than the backward or best subsets methods because the stepwise method is an extension of the forward method with the additional procedure of removing variables previously entered if they no longer contribute to the model. In both of the data sets in which a difference existed between these 2 methods, the forward method gave a larger data set by including a variable that became redundant when later variables were added by both methods.

Backward/stepwise comparison

The backward method differed in a consistent manner from the stepwise method in 2 ways. In each of the 5 data sets in which they differed the

backward method selected a model with more predictors. If an F-to-enter limit of 4.0 had been used, the backward method also would have frequently given a larger number of predictors. Where the same number of predictors were selected but with different combinations, the stepwise method was more efficient, generally having the higher R^2. In 2 of the 5 cases in which they differed the R^2 values were fairly close but for the other 3 the R^2 values were markedly different (.347/.150, .509/.176, and .967/.000) with the backward method selecting the better model in each case. These 3 data sets all had a combination of variables that acted jointly to predict well but none of the variables entered the model individually in the stepwise or forward methods. These data sets illustrate that in certain circumstances the stepwise and forward methods can select very inadequate models.

Backward/best subsets comparison

On 12 of the 16 data sets the same model was selected by the backward and best subsets methods. The worst discrepancy between the models selected by the two methods was on the GMB1 data set in which the models had 5 and 8 predictors and R^2 of .716 and .739. It seems as if the backward and best subsets methods can be counted upon to give models that are reasonably similar in number of predictors and amount of variance explained, although if there is a difference the backward method generally will give a larger model. In the 4 data sets in which the 2 methods gave different models, the backward method selected a larger model 3 times and a smaller model once (although this was due to a tolerance problem).

Stepwise/best subsets comparison

Excluding the 3 cases in which the stepwise method was very inadequate and the case with the tolerance problem, the number of predictors selected by the stepwise method was the same as that selected by the best subsets method in all but 3 cases where the stepwise method gave 1 additional predictor in each case. The additional variable in each of the larger models barely

entered over the F-to-enter of 2.00 level and the discrepancy should not be considered important but more of an indication that the F-to enter level of 2.00 was not exactly equivalent to the criterion of the lowest C_p value that was used in the BMDP9R program.

Best subsets summary

The algorithm used in BMDP9R, which admittedly does not compare all possible models, will not always list all "good" models. In the GMB1 data set, the 8 predictor model chosen by the backward method was not even listed as one of the alternatives in the BMDP9R output even though it had a higher R^2 than all but one of the alternatives that were mentioned. The best subset method, however, does seem to work the best of all of the prediction methods with the data sets used here. It is especially recommended because it encourages a non-mechanical selection process by giving many suggested models.

Backward summary

The backward method can be counted on to give a model which will explain about as much variance as models chosen by any other method but it may include more variables than are necessary to get a "good" model. A major danger occurs with this method, however, when there is high multicollinearity. In this case, computational inaccuracies may occur, so tolerance problems should be considered before running a backward solution.

Stepwise summary

The stepwise method will generally give a model that comes close to maximizing the amount of variance explained for a given number of predictors. If conditions of multicollinearity, suppression, and sets of variables working jointly do not occur, the models selected by the stepwise method can be expected to be as good as the models selected by the backward and best subsets methods. If these conditions do occur, however, the stepwise method may give a model that is completely inadequate. To guard against this occurrence, the

stepwise method should never be used alone to select a model, but only in conjunction with the backward and/or best subsets methods.

Forward summary

The forward method, although discussed in almost all regression textbooks, is rarely, if ever recommended as a reasonable alternative to the stepwise method, and this paper supports the idea that the method has little merit if the stepwise method is available.

Selection process summary

When a model is to be selected, it is important to consider more than one procedure. If one method is to be used, it would appear that the best subsets method is the best of the methods examined here since the computer program generates many models from which a "best" one can be selected. The virtue of running a backward and/or stepwise solution in addition to the best subsets method would be to identify differences in the models that point out characteristics of the variables and/or data set that might be overlooked otherwise. Using the best subsets or backward procedures, it is unlikely that an extremely poor model would be chosen, but this is a real possibility with the stepwise and forward methods. For this reason it is recommended that the stepwise and forward methods NEVER be used alone in selecting a model for any purpose.

Table 1

Regression Models Selected by Different Selection Methods

DATA SET	N	IV's	Number of Predictors Selected/Differences from Stepwise/R^2 Stepwise #	Stepwise R^2	Forward #	Forward +	Forward -	Forward R^2	Backward #	Backward +	Backward -	Backward R^2	Best Subsets #	Best Subsets +	Best Subsets -	Best Subsets R^2
GMA1	49	6	3	.497												
GMA3	13	6	4	.999	5	1	0	.999								
GMA6	50	14	2	.150					7	6	1	.347	7	6	1	.347
GMA8	33	9	4	.886									3	0	1	.877
GMB1	60	14	6	.727					8	4	2	.739	5	0	1	.716
GMB2A	40	8	4	.878												
GMB2B	40	8	1	.176					5	4	0	.509	5	4	0	.509
TAL	505	16	9	.404												
ENR1	579	7	2	.049												
ENR2	579	14	7	.316												
ENR3	579	28	9	.521					11	3	1	.525	8	0	1	.520
ENR4	579	14	5	.089												
ENR5	579	28	14	.361												
LONG	16	6	3	.985									4	1	0	.995
HALD	13	4	2	.979	3	1	0	.982								
SUP	10	3	0	.000					3	3	0	.967	3	3	0	.967

= number of predictors selected using F=2.0 for entry and F=1.99 for deletion for the stepwise, forward and backward models and C_p=2.0 for the best subsets model.

+ = number of predictors selected in this model that were not in the stepwise model

- = number of predictors in the stepwise model that were not selected in this model

References

Berenson, M. L. et al. (1983). Intermediate Statistical Methods and Applications. Englewood Cliffs, New Jersey: Prentice-Hall.

Chatterjee, S. & Price, B. (1977). Regression Analysis by Example. New York: John Wiley & Sons.

Cohen, J. & Cohen, P. (1975). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Dixon, W. J. & Brown, M. B. (1979). BMDP-79 Biomedical Computer Programs P-Series. Berkeley, California: University of California Press.

Draper, N. R. & Smith, H. (1981). Applied Regression Analysis, Second Edition. New York: John Wiley & Sons.

Freund, R. J. & Minton, P. D. (1979). Regression Methods. New York: Marcel Dekker.

Gunst, R. F. & Mason, R. L. (1980). Regression Analysis and its Application. New York: Marcel Dekker.

Kleinbaum, D. G. & Kupper, L. L. (1978). Applied Regression Analysis and Other Multivariable Methods. North Scituate, Massachusetts: Duxbury Press.

Lohnes, P. R. & Cooley, W. W. (1968). Introduction to Statistical Procedures: with Computer Exercises. New York: John Wiley & Sons.

Longley, J. W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. Journal of the American Statistical Association, 62, 819-831.

Lutz, J. G. (1983). A method for constructing data which illustrate three types of suppressor variables. Educational and Psychological Measurement, 43, 373-377.

Morris, J. D. (1975). A computer program to create a population with any desired centroid and covariance matrix. Educational and Psychological Measurement, 35, 707-710.

Morrison, D. F. (1983). Applied Linear Statistical Methods. Englewood Cliffs, New Jersey: Prentice-Hall.

Neter, J. et al. (1983). Applied Linear Regression Models. Homewood, Illinois: Richard D. Irwin.

Pedhazur, E. J. (1982). Multiple Regression in Behavioral Research. New York: Holt, Rinehart, Winston.

Wilkinson, L. (1984). SYSTAT. Evanston, Illinois: SYSTAT, Inc.

Younger, M. S. (1979). A Handbook for Linear Regression. North Scituate, Massachusetts: Duxbury Press.

MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 16, NUMBER 1, SUMMER 1986

Microcomputer Selection of a Predictor Weighting Algorithm

John D. Morris
Florida Atlantic University

An empirical method (PRESS) for examining and contrasting the cross-validated prediction accuracies of some popular algorithms for weighting predictor variables was advanced and examined. The weighting methods that were considered were ordinary least squares, ridge regression, regression on principal components, and regression on an equally weighted composite. PRESS was executed on several data sets having varied characteristics, with each of the weighting techniques obtaining the greatest accuracy under some conditions. The degree of advantage or disadvantage offered by these alternate weighting algorithms relative to ordinary least squares was considered. As it was not possible to determine a priori which weighting technique would be most accurate for a particular data set from theoretical knowledge or from simple sample data characteristics, the sample specific PRESS method was proffered a possibly most appropriate when the researcher wishes to select from among the several alternate predictor weighting algorithms in order to achieve maximum cross-validated prediction accuracy. The feasibility of the use of a microcomputer for the computation intensive PRESS algorithm was also considered.

Many empirical and theoretical studies (Darlington, 1978; Dempster, Schatzoff, and Wermuth, 1977; Gibbons, 1981; Morris, 1979; Pruzek and Frederick, 1978; Wainer, 1976) have suggested that there are more accurate (in the sense of cross-validatior predictor weighting strategies than the traditionally used ordinary least squares (OLS).

Much of the effort has concentrated on ridge regression, with Darlington's (1978) recommendations being by far the strongest in the behavioral sciences. However, some more recent results (Morris, 1982, 1983) suggest a less enthusiastic outlook toward ridge regression in the specific situations considered by Darlington (1978), but a possibly more promising outlook under other data conditions (Morris, 1981). Additional evidence and reservations of others about ridge regression may b found in Egerton and Laycock (1981), Pagel and Lunneborg (1985), Rozeboom (1979), a Smith and Campbell (1980).

Similar controversy spanning at least a quarter of a century (Claudy, 1972; Dawes and Corrigan, 1974; Dorans and Drasgow, 1978; Einhorn and Hogarth, 1975; Gabriel, 1980; Laughlin, 1978; Lawshe and Schucker, 1959; Pruzek and Frederick, 197 Schmidt, 1971; Trattner, 1963; Wainer, 1976, 1978; Wesman and Bennett, 1959) has surrounded the use of equally weighted predictors as a substitute for OLS weights. In addition, several investigators have proposed the use of reduced-rank prediction methods to enhance cross-validation prediction accuracy, possibly beginning with Burkett (1964), to more recently (Morris and Guertin, 1977; Pruzek and Frederick, 1978).

It seems clear that claims for a "panacea" weighting technique to fit all data configurations, such as ridge coefficients "will undoubtedly be closer to (the true parameters) and are more stable for prediction than the least squares coefficients" (Hoerl and Kennard, 1970, p. 72), or "Ridge regression is the best technique for a broad range of intermediate values of validity concentration and is little worse th alternative techniques at the extremes" (Darlington, 1978, p. 1250) are unrealistic Equally clear is that many simulation results strongly suggest that non-OLS weighti strategies offer the researcher enhanced cross-validation prediction accuracy in ma data configurations. The most important next step seems to be to determine the frequency with which such data configurations that are conducive to non-OLS methods occur in the behavioral sciences and to examine the importance of the gain or loss resultant from using these strategies. Given encouraging gains in a reasonable proportion of available data sets, another step would be to generate mechanisms for helping the researcher decide which of the alternate weighting techniques are best for which data situations, and for estimating how much improvement or degradation might be realized by using an alternate technique instead of OLS in a specific data set.

Some simulation results (Morris, 1981, 1982; Pagel and Lunneborg, 1985;) have yielded some general suggestions for when to use which technique. One major factor suggested by Pruzek and Frederick (1978) and explicated more explicitly by Darlingt (1978), is validity concentration, the degree to which predictive validity is concentrated in the first few principal components of the predictors. From simulation results and theory (Darlington, 1978; Morris, 1982; Pagel and Lunneborg, 1985), it is know that as predictor variable collinearity and validity concentratio increase, non-OLS methods usually become more accurate than OLS at some point. In addition, Cattin (1981) has argued that in typical behavioral science data small eigenvalues from the predictor variable intercorrelation matrix tend to explain mor

noise than signal. Thus as the validity concentration is high, non-OLS methods are usually most accurate. However, this tendency is diminished by an opposite trend in favor of OLS regression as sample size and population multiple correlation increase. How these trends balance out with real data is not immediately apparent.

These effects also depend on the type of prediction accuracy of concern. Many simulation studies have concentrated on the error in estimating population regression weights. Instead, the interest in this paper is on the accuracy of criterion score prediction. This accuracy criterion seems more reasonable than that of the accuracy of estimating population regression weights because such techniques as ridge regression may be inappropriate when the sizes of regression weights are of primary concern (Darlington, 1978; Pagel and Lunneborg, 1985). Moreover, the same analytic strategy illustrated in this paper is generalizable to the task of examining errors in estimating regression weights.

However, even when limiting consideration to prediction, one must consider both "relative" and "absolute" types of prediction accuracy. Is the researcher interested in generating a prediction equation that yields predicted scores that are maximally correlated with the actual criterion score (relative), or is the goal to minimize the differences between the actual and predicted criterion scores (absolute)? These are not the same goals, and the comparative accuracies of the methods are partially a function of which one is considered.

Some theoretical (Thisted and Morris, 1980) as well as empirical (Musgrage, Marquette, and Newman, 1982) rules have been offered for determining when various types of ridge regression may be helpful in enhancing prediction accuracy. These rules do not specifically consider the effects either of validity concentration or of sample size, both of which have been shown in simulation studies to affect the relative performance of OLS and non-OLS methods. Also, as operating characteristics for these theoretical rules have not been examined through simulation, it is difficult to know how they would perform with real data. As well, the rules due to Thisted and Morris consider only ridge regression as an alternative to OLS regression.

Although some general trends and suggestions may be gleaned from these studies, it is at best difficult to suggest to an applied researcher what method to select given the specific data characteristics of a sample. The results are useful theoretically, but they are just not sufficiently simple to allow easily applicable rules to be generated to use for specific data sets. Also, such rules would require unknown population information for which one has no sample estimate, as in the case of validity concentration.

More important, very little, if any, information is available about how much gain or loss in prediction accuracy one might expect by using non-OLS weighting with <u>real data</u>. What is the potential payoff or loss for the researcher in trying these non-traditional methods?

Purpose

The purpose of this paper was to advance and examine an empirical sample-based method (PRESS) to be used for exploring the comparative performance of several predictor weighting methods on a specific data set to aid in selection, and most important, to assist in judging the probable resulting gain or loss in prediction accuracy in selecting a weighting algorithm. Although the specific technique is different, the use of an empirical sample-based method to aid in selecting a

predictor weighting method is parallel with the suggestion of Dempster, Schatzoff, and Wermuth (1977, p. 106) that "it would seem that comparison of the predictive capabilities of various methods from one subset to another would provide a reasonable empirical basis for selecting a particular method in a given situation." To demonstrate the technique, the PRESS algorithm was executed on several typical, although not necessarily completely representative, sets of data. The feasibility of the use of a microcomputer for the computation intensive PRESS algorithm was also considered.

The PRESS Algorithm

Allen (1971) introduced a technique that he labeled PRESS (PRedicted Error Sum of Squares) to be used to select a multiple regression variable subset that would yield a minimum sum of squared errors in prediction on cross-validation. This algorithm is executed by alternately predicting each subject's criterion score from the regression equation generated from the predictor and criterion scores of all other subjects. The resulting squared errors of prediction over all subjects are accumulated and the sum obtained serves as a criterion for cross-validation accuracy.

Although most of the multiple regression literature dealing with this "round-robin" subject deletion strategy references Allen and terms the technique PRESS, it is not original with Allen. Perhaps the earliest explicit description of the technique was in a paper by Gollob (1967). Many researchers, however, have recommended the procedure for both multiple regression and discriminant analysis-type classification cross-validation (Allen, 1971; Allen and Cady, 1982; Lachenbruch and Mickey, 1968; Mosteller and Tukey, 1968; Stone, 1974). Additionally, the technique has also been descriptively termed "leave-one-out" (Huberty, 1984; Huberty and Mourad, 1980).

Allen (1971) also provided a derivation for a computational simplification used in calculating PRESS that requires only one matrix inversion, rather than the implied n inversions, where n is the total number of subjects. This derivation was based on a matrix identity often attributed to Bartlett (1951), although no mention was made of Bartlett's work. However, one also can find the same identity in Horst (1963, p. 428) with no mention of Bartlett. Whether all three authors independently derived the same matrix identity is unknown.

Although this algorithm was introduced to help select a subset of predictors that would yield the smallest sum of squared errors upon OLS cross-validation and to give an estimate of the resulting cross-validated prediction accuracy, the same logic and algorithm can be used to judge the cross-validated prediction accuracies (relative or absolute) of alternate predictor weighting methods; the idea is completely general across any weighting strategy. PRESS can be performed for each competing predictor weighting method, and the most accurate method can be chosen as the one most probable to be most accurate on use in replicate samples, or the researcher may decide that the gain, if offered by a non-OLS strategy, is not important enough to warrant selection of a method that may not be well known.

The computational simplification offered by Allen (1971) is rather straightforward for OLS. If one considers the usual model for multiple linear regression,

$$Y = BX + e,$$

where X is an $n \times p$ matrix of $p - 1$ predictor variable values and the usual unit vector, Y is the vector of criterion scores, and e is the vector of error terms, the

al solution for B, the vector of regression weights, is
$= (X'X)^{-1} X'Y$.
deleting a subject would change both X'Y, and X'X, it would seem that both X'Y,
the matrix inverse $(X'X)^{-1}$ would need to be recalculated as each subject is
leted.

However, if $\hat{Y}_{(i)}$ is a subject i's predicted criterion score when that subject's
tor of predictor scores, X_i, and criterion score, Y_i, are excluded from X and Y,
len (1971, p. 11) showed that
$= (1 - Q_i)^{-1}Y_i - Q_i(1 - Q_i)^{-1}Y_i$,
ere $Q_i = X'_i(X'X)^{-1}X_i$, Y_i is the subject's criterion score predicted from the
gression weights based on all the sample, and Y_i is the subject's actual criterion
ore. Although this formulation avoids the numerous matrix inversions, it still
quires the calculation of the predicted criterion score and the Q_is for every
bject. This calculation route, which was found to be as much as an order of
gnitude faster than actually calculating the inverses in a recent comparison
rris, 1984), requires very little extra computation if one ordinarily calculates
siduals.

The most obvious step would then seem to be to try to adapt this computational
rtcut for use with the non-OLS methods of interest. In fact, by recognizing the
lationship between OLS, principal component, and ridge regression, one not only can
opt the algorithm, but also can do the calculations for the methods essentially
multaneously. As well, the Allen formulation obviously fits the case of regression
an equally weighted composite, as regression on such a composite just turns out to
a case of simple regression.

In fact, in a later publication, Allen (1972) provided a version of the shortcut
rmula for ridge regression. Given the usual simple ridge regression model of
$= (X'X + kI)^{-1} X'Y$,
len showed that it followed that PRESS can be calculated from the same formulation
with OLS except that the kI would be added to the X'X matrix before inversion in
e calculation of $\hat{Y}_i$ and Q_i.

However, there is a problem with this formulation. When the researcher decides
on a biasing "k" in ridge regression, it is added to the correlation matrix rather
an to X'X. Although one can center and scale the score vectors such that X'X = R,
e formulation is still incorrect since kI is being added not to the correlation
trix, but to the correlation matrix decreased by the contribution of one subject.
an illustrative problem with five subjects and one predictor variable (X = 2, 0,
3, 9; Y = 3, 4, 4, 7, 6; and a Dempster, Schatzoff, and Wermuth [1977] RIDGEM k
2.73), the PRESS cross-validated correlation calculated by the shortcut formula
as -.92, but the true PRESS cross-validated correlation calculated by actually
nverting n correlation "matrices" augmented by kI was -.07. This example is
rtainly not purported to be representative. Moreover, the difference would clearly
e less for samples of even moderate size and with smaller ks. However, it does
llustrate that the Allen shortcut formulation for ridge regression gives incorrect
esults.

Another difficulty, however, stems from the fact that for ridge regression, the
used is often derived from characteristics of the sample. Thus it is also a random
ariable. As the accuracy of the choice of k affects the accuracy of the resulting
rediction equation, the algorithm for that choice must also be cross-validated.
his task is clearly not accomplished in the Allen shortcut formulation. The same

argument can be advanced for any choice made using information from the data of the sample that affects the prediction equation. Thus one also must cross-validate the algorithm for selecting the number of components in regressing the criterion on principal components, and for choosing the algorithm for deciding which variables are "salient" enough to be included in an equally weighted composite, if such judgments are to be made from sample information.

If one adopts this philosophy of cross-validating the total choice process involved in constructing a prediction model from sample data, then the only computational route possible is to calculate n versions of each equation by actually leaving a subject out each time.

A Pascal computer program was written that cross-validates OLS, ridge regression, regression on principal components, and regression on an equally weighted composite via PRESS for any input data set. One of the difficulties with such techniques as PRESS, bootstrapping (see Efron, 1979; 1983), and other resampling plans is the extreme amount of computation required. When using a mainframe or minicomputer, this translates into costly run times. As microcomputers are a "one-time" expense, such computation costs essentially nothing given the availability of the machine and software. A disadvantage of the microcomputer is that it is slower than mainframes and minicomputers. However, the degree of difference in speed is rapidly decreasing with the continuing introduction of faster and more powerful microprocessors. With this in mind, this program was used with an MS DOS microcomputer to illustrate and to examine the method on several sets of data, and to assess the performance of the microcomputer in accomplishing these relatively demanding computational tasks.

Method

Weighting Techniques

There are many possible choices for a k for ridge regression. Because of its excellent performance and its ease of calculation, the Lawless and Wang (1976) k, which is the inverse of the F ratio resulting from a test of the OLS R^2, was used for ridge regression.

Because of its ubiquity, the Kaiser (1960) rule of selecting components with roots larger than one was used to select the number of components in regressing a criterion variable on principal components. One might also consider using a significance test (e.g., Bartlett, 1950) to determine the number of predictor components to use. One should note, however, that a subjective decision would be necessary even though a significance test is used as the researcher must select a significance level.

As is often the practice, equal weighting was accomplished by specifying a threshold predictor-criterion correlation for inclusion of a predictor. The predictor then received either a +1 or -1 weight depending on the sign of the predictor-criterion correlation. The resultant composite was then used to predict the criterion. For the example data sets presented in this paper, predictor variables with a correlation significant at the .05 level were included.

Obviously, if other non-OLS strategies were used, different results might have been obtained. Likewise, with other data sets, results might have been different. The purpose, however, was a demonstration of a method for examining and comparing the accuracies of the weighting methods for specific data sets rather than a general comparison of the weighting methods.

The Demonstration Data Sets

Twenty-one data sets of widely varying characteristics from the behavioral and natural sciences were used in this demonstration. An attempt at sampling a variety of types of data was made; however, the data sets are not advanced as representative. The results were not intended and should not be interpreted as generalizable to all behavioral science data sets. The intent was to explore and to demonstrate a strategy for estimating what one might expect for a specific data set.

It also is important to note that the actual "real" data sets were used rather than Monte Carlo simulations from covariance structures as has been done in some studies mentioned previously. This procedure not only allows the characteristics of the data structures to vary as they do in nature, but also affords the unique distributional characteristics of a sample to affect the results, contrary to the situation in simulation studies in which multivariate normality is usually assured.

These data sets actually have been used in regression analyses. They are from journal articles, paper presentations, or text books. Therefore any aberrant score vectors are assumed to have been deleted. Before applying the PRESS strategy (or any other analytic method), the researcher probably would wish to consider the removal of "outliers" that manifest appreciable leverage. One may find it helpful to consider the excellent review by Hocking (1983), as well as associated comments for information on methods for detecting such score vectors.

Results

Tables 1 and 2 show the performance of the four weighting techniques for each of the 21 data sets. In concentrating on relative prediction accuracy Table 1 furnishes cross-validated correlations; Table 2 provides absolute accuracy as the mean squared error in predicting the criterion score. In both tables there appears (a) a short description of the origin of each data set (exact citations being available on request), (b) the OLS squared multiple correlation calculated in the total sample (RSQ), (c) the multicollinearity index due to Thisted and Morris (1980) (MI), (d) the ratio of the number of subjects to predictor variables (n/p), and (e) the performance of the methods, with the performance of the non-OLS methods shown as a percent of the OLS performance. It should be noted that the MI criterion proposed by Thisted and Morris is different when one considers relative and absolute accuracy.

The number of subjects ranged from 16 to 293, and the number of predictor variables varied from 3 to 17. The largest raw score matrix analyzed had 271 subjects with 12 predictors.

An interesting characteristic exhibited in the results is the amount of variety obtained. The comparative performance of the methods is clearly dependent on which data set is being considered and on whether the criterion of accuracy of concern is relative or absolute. In addition, the fact that the different methods performed better with differing real data sets may lend some credibility to such differences found in simulated data sets.

Relative Accuracy (Table 1)

Relative accuracy is discussed first. In 16 of the data sets of Table 1 (1, 2, 3, 4, 6, 7, 9, 10, 11, 12, 13, 14, 15, 18, 20, and 21) ridge performance was about the same as that of OLS (within 2%). However, within these same data conditions, the accuracies of regressing the criterion variable on principal components, and of regressing the criterion variable on an equally weighted composite were much less

consistent. Sometimes these procedures were also very close to OLS performance. In one data set (10) they were about 10% better than OLS. Moreover, they ranged down to being appreciably inferior to OLS regression for equal weighting (as evidenced in 11, 13, 14, 18, 20, and 21) to drastically inferior for regression on principal components (6, 13, 14).

Ridge regression was appreciably superior to OLS regression in relative accuracy on four data sets (8, 16, 17 and 19) ranging from 11% up to 44% better than OLS regression. However, for all these four data sets, at least one (in two cases both) of the other non-OLS methods were considerably superior to ridge - - an outcome much like that provided by the results reported in a previous simulation study of relative accuracy (Morris, 1982).

In one data set (5), ridge did very poorly on relative prediction accuracy, as evidenced by yielding a negative cross-validated correlation (as principal components did in data set 16). Yet regressing the criterion variable on principal components or on an equally weighted composite performed much better than OLS. However, the importance of this particular result must be viewed in context; even though the squared multiple correlation was an appreciable .817, the cross validated OLS correlation was only .028 so that no meaningful prediction could take place on replicate samples in any case.

Absolute Accuracy (Table 2)

As for absolute accuracy (Table 2), the results were different. In 12 of the data sets, (1, 2, 3, 4, 6, 10, 12, 13, 14, 15, 20, and 21) ridge was within about two percent of the mean squared error produced by OLS regression. (It should be noted that smaller is superior for this measure of accuracy.) These data sets constituted a subset of the 16 meeting this same criterion for relative accuracy. On these same 12 data sets regressing the criterion variable on an equally weighted composite followed the results of ridge fairly closely; although superior (ranging from very slightly to appreciably) to ridge regression on three data sets (3, 10 and 12) regressing on an equally weighted composite was inferior on the rest. Regressing the criterion variable on principal components displayed much more variety within these 12 data sets. Performance was about the same as ridge on five of the data sets (2, 3, 15, 20, and 21), superior on three data sets (1, 10 and 12), and ranged to drastically inferior (4, 6, 13, and 14).

On eight of the data sets (5, 7, 8, 9, 16, 17, 18, and 19) ridge was appreciably better than OLS regression in absolute accuracy, with the decrease in mean squared error of prediction ranging from about 4% (data set 18) up to nearly 70% (data set 5). In four (5, 8, 16, and 19) of these eight data sets both regressing the criterion variable on principal components and on an equally weighted composite were in turn considerably better than ridge.

In only one data set (11) did ridge not perform at least about as well as OLS on absolute accuracy, with a mean squared error of about 21% more than that for OLS regression. Both principal components and equal weighting also performed very poorly on this data set. It is quite interesting and possibly important to note that this is not the same data set as the one on which ridge was so poor in relative accuracy; on that data set (5), ridge exhibited its best absolute accuracy performance (only 31% of the mean squared error of OLS regression).

Although the results from this data set may need to be considered especially cautiously because of the very small cross-validated correlation, the results also

did not agree between relative and absolute accuracy in other instances. The decision of whether one is primarily interested in relative or absolute accuracy is an important one.

For these data sets, the number of subjects per variable, multicollinearity, and sample OLS multiple correlation all appeared to be of no use in helping the researcher decide whether one of the non-OLS methods would be worth pursuing. The question of identifying the most accurate prediction method is really one of classification. Can one "classify" a data set to the method yielding the greatest accuracy from sample characteristics? Using the "leave-one-out" strategy of Lachenbruch and Mickey (1968), these three sample characteristics were unable to classify the data sets into the most accurate strategy (OLS or non-OLS) any better than chance assignment would have for both relative and absolute accuracy. In fact, when combining the results for both relative and absolute accuracy, the number of correct classifications was exactly the same as one would expect by chance. For this reason, it would not seem possible to construct rules for deciding a priori from these statistics arising from a specific sample which method would be likely to be most accurate on application to a replicate sample.

Discussion

Any summative comments that could be made related to the relative performance of the methods are necessarily only relevant to these data sets. Moreover, the purpose of this study was not to declare a best method, or even to derive rules based on sample characteristics for deciding which strategy to use. Indeed, the inability to explain easily the behavior of the weighting techniques from the sample characteristics presented argues for just such a sample specific approach as has been used and is being proffered.

One generalization that probably can be made from the results, however, is that none of the non-OLS methods offers a panacea for achieving maximum accuracy across all data sets as some reports in the literature might suggest. The researcher stands to lose a lot of prediction accuracy by choosing any of the non-OLS strategies under some data conditions. Likewise, the researcher stands to gain a great deal in some data conditions if a superior algorithm can be selected. The problem is that it is not easy to specify under what circumstances the realization of a superior algorithm will occur from simple sample data characteristics; thus, the more complicated PRESS procedure may be called for.

Although the data sets utilized in this paper may not be representative, it may still be reasonable to suggest that the performance of none of the non-OLS methods was good enough often enough to recommend routine application of them in the same way that OLS regression is used. At the same time, moreover, there are appreciable accuracy gains possible in some cases. If prediction accuracy is sufficiently important for the data set and situation at hand, the researcher may wish to take the trouble to ferret out those occasions for which a more accurate non-OLS procedure can deliver greater accuracy; the PRESS algorithm is suggested as a viable strategy for that task.

The computation times for all the runs are included in Table 3. Most of the runs only took a few seconds, with several taking a few minutes. The two largest jobs in which the Project Talent data was analyzed separately by sex each took more than an hour to run. Whether the times are reasonable or not is clearly a subjective decision. However, even times of more than an hour don't compare unfavorably with

the batch job turn-around time that can be expected when using many large computers.

The microcomputer used was a Sanyo MBC 550. This is an MS DOS machine with an 8088 microprocessor. It is similar in many ways to an IBM PC, but the 8088 clock rate is slower (3.6) than that of the IBM PC (4.77). An 8087 arithmetic coprocessor was also installed to aid in speed and accuracy. Because of the slower clock rate, almost all IBM PC "clones" would run these jobs faster than the times represented.

The computer language used was Turbo Pascal. While a good performer in general, it is certainly not the fastest "number crunching" language available. For example, a recent article in BYTE found the Microsoft Pascal compiler to run a computation intensive program utilizing the 8087 nearly twice as fast as Turbo Pascal. Microsoft Pascal, however, was unavailable to test. The Pascal program should run with no modification.

It should also be noted that newer, faster, and more powerful microprocessors are now commonplace. The 8086, 80186, and the 80286 of the IBM AT should all perform better than the times represented here. Therefore, for all these reasons, the times presented should be considered as quite conservative. Moreover, a 32 bit 80386 has recently been released and will be much faster (probably by a factor of more than four) than the fastest of these (the 80286). Super microcomputers with the power of a VAX mini should be on our desks very soon.

While microcomputer time is essentially free, a deficit in a long running job is that the machine is generally lost for other uses. However, there are now some good multitasking systems available that will allow the use of the computer for other purposes, i.e. word processing, while such a computation laden job is number-crunching in the "background." Such multitasking systems will almost certainly be a standard part of the operating system of the more powerful microcomputers that will be common in the very near future.

Although several strategies can be employed to make the computing algorithm as efficient as possible, a large amount of computation may result in any case. In general, in judging whether the PRESS technique is worth pursuing a researcher would need to consider the size of the prediction problem and resulting costs of PRESS in relation to the relative importance of the goal of maximizing prediction accuracy. It is important to note, however, that most prediction problems seen in the behavioral science literature are not excessively large and that in any case the non-OLS methods are really only contenders with relatively small samples. Further, the trend of the decreasing cost of computational power is accelerating; researchers need to plan their methods such that they can capitalize on this resource. Tukey's (1985) comments relating to our need to make sure that the statistical techniques we invent anticipate the incredible resources of computational power that we will have in the near future seem especially relevant.

A copy of the Pascal computer program is available for those wishing it. It is a COM file and should work on any MS DOS microcomputer with a microprocessor in the Intel 8088, 86, 286, etc. line. In requesting the program, please specify whether the program can expect to find an 8087 arithmetic processing unit available. If the program is of interest, send a blank DSDD diskette to:

John D. Morris
College of Education - IRDTE
Florida Atlantic University
Boca Raton, Florida 33431

References

Allen, D. M. (1971). The prediction sum of squares as a criterion for selecting predictor variables (Tech. Rep. No. 23). University of Kentucky, Department of Statistics.

Allen, D. M. (1972). Biased prediction using multiple linear regression (Tech. Rep. No. 36). University of Kentucky, Department of Statistics.

Allen, D. M., and Cady, F. B. (1982). Analyzing experimental data by regression. Belmont, CA: Wadsworth.

Bartlett, M. S. (1950). Tests of significance in factor analysis. British Journal of Psychology, 3, 77-85.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. Annals of Mathematical Statistics, 22, 107-111.

Burkett, G. R. (1964). A study of reduced rank models for multiple prediction. Psychometric Monographs, No 12.

Cattin, P. (1981). The predictive power of ridge regression: Some quasi-simulation results. Journal of Applied Psychology, 66, 282-290.

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

Darlington, R. B. (1978). Reduced-variance regression. Psychological Bulletin, 85, 1238-1255.

Dawes, R. M., and Corrigan, B. (1974). Linear models in decision making. Psychological Bulletin, 81, 95-106.

Dempster, A. P., Schatzoff, M., and Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. Journal of the American Statistical Association, 72, 77-91.

Dorans, N., and Drasgow, F. (1978). Alternate weighting schemes for linear prediction. Organizational Behavior and Human Performance, 21, 316-345.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. Annals of Statistics, 7, 1-26.

Efron, B. (1983). Estimating the error of a prediction rule: Improvement on cross-validation. Journal of the American Statistical Association, 78, 316-331.

Egerton, M. F., and Laycock, P. J. (1981). Some criticisms of stochastic shrinkage and ridge regression, with counterexamples. Technometrics, 23, 155-159.

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

Gabriel, R. M. (1980, September). Using composite variables in multivariate analysis: Why weight? Paper presented at the meeting of the American Psychological Association, Montreal, Quebec, Canada.

Gibbons, D. G. (1981). A simulation study of some ridge estimators. Journal of the American Statistical Association, 76, 131-139.

Gollob, H. F. (1967). Cross-validation using samples of size one. Paper presented at the meeting of the American Psychological Association, Washington, D.C.

Hocking, R. R. (1983). Developments in linear regression methodology: 1959-1982. Technometrics, 25, 219-230.

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

Horst, P. (1963). Matrix algebra for social scientists. New York: Holt, Rinehart, and Winston.

Huberty, C. J. (1984). Issues in the use and interpretation of discriminant analysi Psychological Bulletin, 95, 156-171.

Huberty, C. J., and Mourad, S. A. (1980). Estimation in multiple correlation/prediction. Educational and Psychological Measurement, 40, 101-112

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141-151.

Lachenbruch, P. A., and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. Technometrics, 10, 1-11.

Laughlin. J. E. (1978). Comment on "Estimating coefficients in linear models: It don't make no nevermind." Psychological Bulletin, 85, 247-253.

Lawless, J. F., and Wang, P. (1976). A simulation study of ridge and other regressi estimators. Communications in Statistics - Theory and Methods, A5(4), 307-323.

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

Morris, J. D. (1979). A comparison of regression prediction accuracy on several typ of factor scores. American Educational Research Journal, 16, 17-24.

Morris, J. D. (1981, April). An extension to "A comparison of regression accuracy o several types of factor scores." Paper presented at the meeting of the America Educational Research Association, Los Angeles, CA.

Morris, J. D. (1982). Ridge regression and some alternate weighting techniques: A comment on Darlington. Psychological Bulletin, 91, 203-210.

Morris, J. D. (1983). Stepwise ridge regression: A computational clarification. Psychological Bulletin, 94, 363-366.

Morris, J. D. (1984). Cross-validation with Gollob's estimator: A computational simplification. Educational and Psychological Measurement, 44, 151-154.

Morris, J. D., and Guertin, W. H. (1977). The superiority of factor scores as predictors. Journal of Experimental Education, 45, 41-44.

Mosteller, F., and Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), Handbook of Social Psychology: Vol. 2. (pp. 80-203). Reading Mass.: Addison-Wesley.

Musgrage, B., Marquette, J., and Newman, I. (1982). On using the average intercorrelation among predictor variables and eigenvalue orientation to choos a regression solution. Multiple Linear Regression Viewpoints, 11, 1-21.

Pagel, M. D., and Lunneborg, C. E. (1985). Empirical evaluation of ridge regressior Psychological Bulletin, 97, 342-355.

Pruzek, R. M., and Frederick, B. C. (1978). Weighting predictors in linear models: Alternatives to least squares and limitations of equal weights. Psychological Bulletin, 85, 254-266.

Rozeboom, W. W. (1979). Ridge regression: Bonanza or beguilement? Psychological Bulletin, 86, 242-249.

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

Smith, G., and Campbell, F. (1980). A critique of some ridge regression methods. Journal of the American Statistical Association, 75, 74-81.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions Journal of the Royal Statistical Society, Series B, 36, 111-133.

isted, R. A., and Morris, C. N. (1980). Theoretical results for adaptive ordinary ridge regression estimators (Tech. Rep. No. 94). Chicago: University of Chicago.
attner, M. H. (1963). Comparison of three methods for assembling aptitude test batteries. Personnel Psychology, 16, 221-232.
key, J. W. (1985). Comment. The American Statistician, 39, 12-14.
iner, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. Psychological Bulletin, 83, 213-217.
iner, H. (1978). On the sensitivity of regression and regressors. Psychological Bulletin, 1978, 85, 267-273.
sman, A. G., and Bennett, G. K. (1959). Multiple regression vs. simple addition of scores in prediction of college grades. Educational and Psychological Measurement, 19, 243-246.

Table 1

Weighting Methods' Relative Performance (Cross-Validated Correlation) for Several Data Sets

Numerical Designator and Data Set Description	RSQ	MI	Method			
			OLS	As a % of OLS Ridge	PC	Equal
1 Marquardt's Acetylene Data	.920	1.0	.920	100.01	102.05	99.71
2 Chew LP(5) Predicts MRT	.591	1.0	.750	100.64	100.60	100.30
3 Hoerl's Kansas Corn Yield	.800	1.4	.854	100.24	100.23	100.32
4 Draper and Smith (p. 204)	.914	1.1	.927	100.03	92.67	99.52
5 Drehmer Data (EPM)	.817	1.1	.028	-192.43	379.01	131.28
6 Golf score from Task Perf	.848	1.6	.912	99.99	47.40	99.98
7 Hald Data (D & S, p. 366)	.982	1.0	.980	100.32	99.11	100.27
8 Hocking & Dunn RR Symp. '82	.620	1.0	.318	132.67	230.95	230.59
9 Hoerl RR-1980 Paper	.986	1.1	.979	100.21	100.17	100.17
10 Kerlinger and Pedhazur, 292	.640	1.6	.690	100.46	109.69	109.80
11 Longley D&W p. 312	.996	1.0	.992	99.83	95.61	92.57
12 Journal of Exp. Education	.475	1.1	.635	101.27	104.27	104.42
13 Rulon: Pref & Success - Mech	.261	1.9	.441	98.59	26.49	92.62
14 Rulon: Pref & Success - Oca	.323	2.4	.494	98.68	28.51	74.81
15 Rulon: Pref & Success - Pas	.252	1.5	.432	99.01	94.61	97.25
16 Retention from Demos & WISC	.388	1.2	.058	144.11	-40.94	520.58
17 Piers-Harris from IQ & Ach	.185	2.2	.108	111.96	61.88	142.59
18 D & S Steam Data (p.352)	.949	1.1	.925	99.06	89.44	86.92
19 D & S Data (p. 233)	.816	1.2	.691	111.05	121.17	118.50
20 Female Talent Data C&L p. 345	.331	1.9	.520	100.58	97.78	88.15
21 Male Talent Data C&L p. 349	.411	1.7	.577	101.17	100.54	94.97

Note. Additional information about data sources is available from the author; the abbreviated headings at the top of each data column are described in the text at the beginning of the section concerned with results.

e 2

hting Methods' Absolute Performance (Mean Squared
r) for Several Data Sets

merical Designator and ata Set Description	RSQ	MI	Method OLS	As a % of OLS Ridge	PC	Equal
Marquardt's Acetylene Data	.920	1.1	21.0	99.02	75.60	101.32
Chew LP(5) Predicts MRT	.591	1.0	63.9	98.25	98.25	99.10
Boerl's Kansas Corn Yield	.800	2.1	14.2	98.27	98.37	97.78
Draper and Smith (p. 204)	.914	1.4	13.9	99.16	186.42	106.79
Drehmer Data (EPM)	.817	1.5	1.12	31.20	21.20	26.51
Golf score from Task Perf	.848	2.3	1.98	100.06	485.98	100.21
Hald Data (D&S, p. 366)	.982	1.0	8.49	83.49	141.20	106.61
Hocking & Dunn RR Symp. '82	.620	1.1	53.7	72.57	31.25	31.38
Hoerl RR-1980 Paper	.986	1.4	2.98	90.31	92.58	92.62
Kerlinger and Pedhazur, p.292	.640	2.2	.19	97.63	79.39	79.29
Longley D&W p. 312	.996	1.2	.18E+6	121.39	641.15	1066.6
Journal of Exp. Education	.475	1.4	9.89	97.89	93.62	93.41
Rulon: Pref & Success - Mech	.261	2.4	2.42	100.09	123.62	103.65
Rulon: Pref & Success - Occ	.323	2.7	2.65	100.21	130.28	117.00
Rulon: Pref & Success - Pas	.252	2.0	309.7	99.89	102.17	100.86
Retention from Demos & WISC	.388	1.7	.18E+5	84.15	81.97	58.07
Piers-Harris from IQ & Ach	.185	3.4	209.9	92.12	93.89	93.33
D & S Steam Data (p.352)	.949	1.5	.43	94.83	190.84	208.86
D & S Data (p. 233)	.816	1.7	.007	68.58	48.35	53.23
Female Talent Data C&L p. 345	.331	3.7	2.10	99.16	101.20	107.77
Male Talent Data C&L p. 349	.411	3.2	1.63	98.31	98.79	104.17

Note. The information presented in the Note of Table 1 is appropriate for this table.

Table 3

Score Matrix Size and Computation Times for Several Data Sets

Numerical Designator and Data Set Description	n	p	Time(M:S)
1 Marquardt's Acetylene Data	16	3	:06
2 Chew LP(5) Predicts MRT	293	5	6:51
3 Hoerl's Kansas Corn Yield	51	6	2:02
4 Draper and Smith (p. 204)	21	3	:08
5 Drehmer Data (EPM)	14	9	2:03
6 Golf score from Task Perf	120	4	1:51
7 Hald Data (D&S, p. 366)	13	4	:11
8 Hocking & Dunn RR Symp. '82	20	3	:07
9 Hoerl RR-1980 Paper	15	5	:21
10 Kerlinger and Pedhazur, p.292	30	4	:23
11 Longley D&W p. 312	16	6	:36
12 Journal of Exp. Education	83	4	:58
13 Rulon: Pref & Success - Mech	93	3	:38
14 Rulon: Pref & Success - Oca	66	3	:24
15 Rulon: Pref & Success - Pas	86	3	:33
16 Retention from Demos & WISC	29	10	1:30
17 Piers-Harris from IQ & Ach	55	7	3:40
18 D & S Steam Data (p.352)	25	9	3:29
19 D & S Data (p. 233)	16	4	:13
20 Female Talent Data C&L p. 345	271	12	93:39
21 Male Talent Data C&L p. 349	234	12	80:09

Note. The information presented in the Note of Table 1 is appropriate for this table.

.TIPLE LINEAR REGRESSION VIEWPOINTS
UME 15, NUMBER 1, SUMMER 1986

Discussion of AERA 1986 Session 21.25
Applications of Multiple Linear Regression

Bruce G. Rogers
University of Northern Iowa

Comments on the Paper by Joe Ward:

Since I have not seen the full paper, I will need to base my comments on the short draft I received. It proved to be an innovating application of both a utilitarian philosophical viewpoint and interaction in a simple 2-way ANOVA.

The model is based upon the criteria of maximizing the learning when summed across all students. This is reminiscent of one of the 19th century philosophical discussions on ethics. Jeremy Bentham (b 1748) developed the concept that the criteria of the goodness of a policy was determined by calculating the good for each individual and then summing up the individual goods. Sometimes it is called the calculus approach, in reference to integration as the summing of the values. And that is what is done in the table entitled "Optimality Index Values." For every possible way of assigning the four students to the four teachers, the sum of the Optimality values is computed. Then that particular assignment of pupils with teachers which yielded the maximum sum is chosen as the desired assignment.

Bentham was aware that sometimes the principle of the "greatest good for the greatest number," when applied to public policy, could come in conflict with what a particular individual perceived as their own greatest good. I told Joe that many principals might be hesitant about applying this model for fear of confronting irate parents who wanted another choice. For example, if the parents see the table of Predicted Values, it is likely that

all of them will request that their child be placed with Teacher 1. Trying to convince any one parent to allow their child to be put with a less-than-best teacher in order to maximize some abstract "Optimality Index" may prove to be very challenging. Indeed, it is my understanding that many principals randomly assign pupils to the teachers, when several teachers are teaching the same grade, in order to avoid possible charges of favoritism toward teachers and pupils. But Joe assured me that in some districts (including the one in which his wife taught) the principal and the teachers do consult on how to best assign the students. Given that such decisions are to be made, the Ward procedure has the definite virtue of providing an unbiased approach.

The procedure uses a two-way ANOVA interaction design. It is a variation of the aptitude-treatment interaction, where aptitude is past performance and treatment is the teacher. Richard Snow, Lee Cronbach, and others, have worked extensively to find such interactions, with limited success. However, since the teacher is such an important variable in the classroom, it is possible that this approach will prove to be an efficient method of detecting such interactions.

I like the term "catalytic" variable. In chemistry, we take two compounds which react very slowly or not at all. However, when we add a catalyst, the reaction is speeded up, but the catalyst is not affected. In Figure 1, only a weak interaction is present, but when the catalytic variable is added, a strong interaction is observed, as seen in Figure 2. And the resulting "Optimal Sum of Payoff Values" is increased fourfold, as a result of this interaction.

Let me conclude by making a practical suggestion to the authors. Special computer programs were written to compute the tables. Is it

ossible to do this with regular routines in MINITAB, SPSS, SAS, BMDP, etc.? f so, it would be useful to describe how that is done, thus making the procedures easily available to a large number of readers.

Comments on the papers by Jerome Thayer

In the paper on Model Building, attention is given to a set of widely used approaches to variable selection in multiple regression. It is pointed out that no technique should be used indiscriminantly, but rather, that user judgment should be used to determine that set of predictor variables which will be most interpretable.

These techniques were applied to a variety of data sets, ranging from real world data to contrived data. The results in Table 1 suggest that, in general, the Stepwise method is a desirable procedure, but that exceptions do exist. Therefore, the general consensus does seem to support the author's conclusions.

A suggestion might be made for this paper. The "Best Subsets" program was obtained from BMDP, but is not available in SPSS. What are users to do if only SPSS is available to them? A look at Figure 1 suggests that if the Stepwise and Backward procedures were run, and the highest R^2 selected, the results would not be substantially different from using the Best Subsets procedure. While this point is implied in the paper, perhaps it could be made more explicit.

Thayer's paper on Dichotomous Variables shows an empirical example of the mathematical equivalency of several least squares statistics. The paper first points out that a number of writers in the behavioral sciences have argued that regression is inappropriate for data in which the dependent variable is dichotomous. Thayer chose not to attack the critics directly,

but used that well-known proof model from geometry, reductio ad absurdum. A set of data is analyzed twice, using the dichotomous variable first as the dependent variable and then as the independent variable. The results are shown to be identical. It is then concluded that if the reasoning of the critics was followed to its logical conclusion, it would be necessary to discard t-test, ANOVA, ANCOVA, discriminant analysis, and multiple regression. It would be interesting to hear how the critics would respond to this argument.

But let me suggest a reason why one might prefer a computer program specifically written for each of the above routines, rather than using a regression program only. While it is possible to show that, on a two-group comparison, the t-test, F-test, and simple correlation are mathematically equivalent, the computer output for each is not in the same form. Thus, the square root of F must be taken to get t, and a more complicated transformation must be made to get r to t. It is also true that a 2 group discriminant analysis is the same as multiple regression on a dichotomous dependent variable, but again the computer output looks different. And for more than two groups, the output is much different. If the transformations are not made correctly, then serious differences can result. While that is not the situation that the critics had in mind, it is a legitimate reason why a person might use a technique other than regression.

But I digress. This does not detract from Thayer's basic conclusion that the underlying theory of the various least squares techniques is the same, and therefore all of them can be considered as special cases of multiple regression, canonical correlation, or multivariate analysis of variance (SPSSX uses the latter procedure as an umbrella). Conceptually,

this is a powerful tool for helping the student to see classical statistics as variations on a major theme rather than as a "bag of tricks."

My only suggestion for this paper is that the layout of the tables and the use of the t values may prove difficult for the reader to follow. Perhaps the author will submit the paper to a colleague or a student, and if they have similar difficulties, revise the layout to strengthen the presentation.

Comments on the paper by John Morris

The Morris paper begins by stating that the primary concern in regression is the predicting of accurate criterion scores, rather than the estimating of population regression weights. While it is true that, in the theoretical sense, these two criteria are comparable (i.e., you cannot have accurate criterion prediction without accurate regression weights), it is also true that the beta weights may change if a different type of regression is used (e.g., ridge regression). But in both cases, the ultimate focus is upon the accuracy of the criterion scores.

The PRESS Algorithm was designed to select a multiple regression model variable subset that would minimize the Sum of Squares on Cross Validation. This is somewhat akin to the "best set" selection of which Thayer spoke. The philosophy of cross validating the total choice process (p. 13) by omitting one subject at a time is akin to the "Jackknife" procedure.

In the computer runs, "real" data was used instead of data from Monte Carlo simulations. That definitely has the advantages that are mentioned (p. 15) but also has the disadvantage that one does not know a priori which assumptions are violated and why, whereas with Monte Carlo data we can specify and create the violations. Perhaps in a revision of this paper it

would be useful to discuss both the strengths and weaknesses of these procedures.

The results show that, for most cases, the OLS is sufficient and even better than the other methods. I like this conclusion. It is compatible with my own philosophy of techniques. Some people complain that we use statistics without carefully analyzing the data to see if it meets all the assumptions. But I suggest that if the data even vaguely looks appropriate, we can submit it for computer analysis. Thus, we can examine the results. Do they make sense? If not, what violations might account for it? And how might the data be transformed or the procedure modified to make better interpretable results? The results of this study seem consistent with that. Ridge regression and the techniques have an important place, but for most data we should first look at OLS, and then try other techniques where appropriate. The PRESS algorithm, available on a microcomputer, can then provide an effective way to address this selection problem.

Regression and Model C for Evaluation

Gail Smith, Keith McNeil and Napoleon Mitchell
Dallas Independent School District
Dallas, Texas

OVERVIEW OF SYMPOSIUM

The objectives of this symposium are to:

1) Provide a rationale for using regression analysis (specifically Model C) to evaluate educational programs.
2) Provide one example of an extensive Model C evaluation report.
3) Discuss assumptions of Model C and ways to deal with those assumptions.
4) Share examples of disseminating Model C results to decision makers.
5) Identify and resolve additional technical issues that evaluators need to be concerned about when implementing Model C.

We ask you to pretend that this is the Dallas Independent School District Board meeting. The program manager and evaluator are presenting the end of year evaluation results for a state-funded compensatory education program. The evaluator has briefed the program manager and the report was delivered to the school board approximately two weeks ago. We must assume, though, that no members thoroughly understand the report, mainly because most have not read it in anticipation of being briefed.

The presentation will be made by two evaluators from DISD. Gail Smith will be playing the role of program manager in presenting the basic program. Keith McNeil will be playing the role of evaluator in presenting the evaluation results. A third evaluator from DISD, Napoleon Mitchell, will be playing the role of court-appointed auditor, questioning the precedures, results, and interpretations. (Those of you who do not have the pleasure of working under the constraints of a court order may want to think of Napoleon as a board member who has a Ph.D. in statistics and doesn't mind you knowing it.) We would appreciate you asking your questions only after the auditor is satisfied that all his questions have been asked/answered. The last ten minutes of the symposium is reserved for the comments from our distinguished discussant, Dr. George Powell of the Educational Testing Service.

DESCRIPTION OF TREATMENT PROGRAM

The goal of the Reading Improvement Program was to narrow the gap in reading performance between lower and higher achieving students as well as minority and White students in the District. Objectives for accomplishing this goal included: a) providing an additional two-semester, reading course with a restricted teacher pupil ratio of 1:20, b) providing special curriculum materials in logic, vocabulary, comprehension, and study skills, and c) providing staff development on effective instructional strategies in reading to participating teachers. The additional language arts course, focusing on reading, was required for students in grades seven and eight who scored below the 40th percentile in Reading Comprehension on the Iowa Tests of Basic Skills (ITBS). All students scoring below the 40th percentile at all 24 District Middle Schools were eligible for the program with two exceptions. Students in special education classes and students in the two beginning levels of English-as-a-Second-Language classes were not eligible.

Characteristics of students enrolled in the program are presented in Tables 1 and 2. The figures in Table 1 indicate that nearly half the

Table 1

Number of Students Enrolled
and Not Enrolled in RI Course
Fall, 1984

Enrolled in RI Course	Grade 7	Grade 8
Yes	4285	4790
No	5374	4383
Total	9659	9173
% of Total in RI Course	44	52

students in the District middle schools were enrolled in the program in the fall of the second year. The analysis of program effectiveness was conducted using ITBS reading comprehension test scores for both Spring 1984 (pretest) and Spring 1985 (posttest). The number of students represented in this analysis is provided by race and grade in Table 2. Ethnic minority students comprised 87% of the total number of participating students at both grades seven and eight.

Table 2

Grade	Stat	Ethnicity Black	Hispanic	Asian/Indian	White	Total
7	N	2111	591	36	398	3136
	%	67.3	18.8	1.1	12.7	
8	N	2580	705	20	482	3787
	%	68.1	18.6	0.5	12.7	

Since the districtwide percentage of minority students was 76%, the RI program was focusing on minority students.

IMPLEMENTATION FINDINGS

The RI program in grades seven and eight was implemented much better than last year, though there were improvements needed. The lack of a program manager with clear lines of authority resulted in lack of communication and slow or erroneous implementation. Staff development sessions were less than successful because of redundancy of topics and timing of material.

Almost all of the classrooms observed appeared to be conducive to learning, although some did have an enrollment of more than the maximum of 20 allowed by the guidelines. Teachers were using the RI texts and support materials, but few were using teaching techniques considered beneficial for these kinds of students.

Few interactions were initiated by students with the teacher controlling the interactions. Although most teachers provided positive reinforcement, not all teachers provided at least five instances of positive reinforcement. The instructional climate was judged to be better in the RI classes than in the regular language arts classes, both in terms of how well the instructional time was used and whether the instruction was conducive to learning.

ACHIEVEMENT FINDINGS

Results for Grade Seven. A total of 3135 RI students had both pre and post scores, although the scores of 151 of these students were eliminated because their post score was considered too deviant in respect

to gains which were either too high or too low to meet normal expectations. Students in RI gained from 30.1 to 32.3 NCE units. But since RI students were selected into the program according to their pretest scores, we would expect the regression effect to raise their scores. RI students also gained more than the comparison group whose pretest scores were above the 40th percentile (2.3 mean gain vs. -5.9 mean gain for the comparison group). Again, though, the regression effect would have predicted the general trend of these results, i.e. the initially higher scoring comparison group showed mean losses while the initially lower scoring RI students showed mean gains.

A significant second degree fit to the data was discovered in the seventh grade comparison group. Hence the Model C analysis employed a second degree curved line of best fit. the curved line of best fit was the same for both the comparison group and the RI group, hence for these eighth grade students there was no effect due to participation in the RI program (See Figure 1).

Results for Grade Eight. A total of 3787 RI students had both pre and post scores, although the scores of 184 of these students were eliminated because their post score was considered too deviant in terms of expected gains or losses. RI students gained from 30.2 to 34.8 NCE units. But since RI students were selected into the program according to their pretest scores, we would expect the regression effect to raise their scores. The RI students gained more than the above 40th percentile comparison students, but again the regression effect would have predicted this outcome.

There was no second degree curvilinear fit found in the eighth grade data, so only linear trends were investigated. Since a significant

interaction was found, an overall program effect was not investigated. The analysis was concluded with the findings of a significant apptitude-treatment interaction. The lines of best fit for the eighth grade are depicted in Figure 2. The RI program is most effective for those students who have the lowest pretest scores. Those students at the program cutoff gain very little from the extra RI class.

ALTERNATIVE EVALUATION MODELS

There are three major ways we could have evaluated this program. These three ways were documented and described by Tahorst, Lmadge and Wood in 1975.

First, we could have compared the performance of DISD children with what we would expect them to do if they were like the national norm group. This has been referred to as the Model A approach, wherein we use the pretest achievement level as the expectation for the posttest performance.

Two major assumptions in the use of this model cannot be met. The selection of students into the program should be independent of the pretest score, otherwise simple regression to the mean can account for substantial movement to the total group's mean. This was the situation in the RI program, as the pretest measure also served as selection into the program.

The second assumption of Model A which cannot be verified is that the students in the norming sample who are at the same pretest percentile levels are like those being evaluated -- like in the sense of demographics and in terms of quality of regular educational curricula. We know that most of the DISD students are inner city students, with a high concentration of low SES students. Therefore, we can't assume that our students are like the national norming sample. The test that we use does have large city norms. Although DISD students consistently score high, we cannot determine if our students gain more than other large city students. The high scores may only reflect higher initial achievement levels of our students. That is, the question of the quality of a program demands assessment of student growth.

Second, we could have used was a local comparison group to evaluate the RI program. This type of evaluation is referred to as Model B in the literature. Model B is difficult to implement in most educational settings, as in this one, because the model requires that some students (who are otherwise qualified) not receive the special treatment. All students scoring below the cutoff of the 40th percentile were supposed to receive the treatment, thus leaving no students for the comparison group. What actually happened was some students below the 40th percentile did not receive the RI course. Some of these students were in special education classes and some received the RI course only one semester. The remaining students did not receive RI for undocumented reasons. It was our educational guess that many of these students were not enrolled in the RI course for educational reasons which would indicate a higher posttest level than indicated by their pretest (e.g. student is really a high achiever, she just didn't pretest well).

The third and final model utilizes a local comparison group which is acknowledgely different at pretest time. The model capitalizes on the fact that this local comparison group receives the same regular curriculum. The expected posttest performance of the treatment group (RI students) is estimated from the performance of the comparison group. This model assumes that the achievement gain is consistent across pretest levels. One of the major problems of Model C is the determination of this consistent trend in achievement gain. Is the trend linear or of some other nature? Another problem is that the presence of erroneous outliers can unduly affect calculations of this trend. Outliers do not affect the calculations of statistics in other models as much as in Model C.

Exhibit 1. Summary of Models

Model Name	Comparison	Expected Post Performance	Problems	Advantages
Model A	students as their own comparison	pretest level	1. selection on pretest 2. students in norming sample—ethnicity, size, quality of program	1. easy to compute by hand 2. similar to what was done in past
Model B	local students who do not receive treatment	posttest of comparison students	1. students denied service 2. requires testing of additional students	1. both groups of students receive similar regular curriculum
Model C	local students who do not receive treatment	predicted from comparison students	1. linearity 2. outliers 3. calculation and interpretation	1. both groups of students receive similar regular curriculum 2. don't have to deny services to some students 3. can test for aptitude by treatment interaction 4. can reflect non-linear reality

Model C was chosen as the best model to evaluate the program because students were selected into the program on the basis of their pretest scores, and most students below the cutoff score were served. Those that were not served did not constitute a valid comparison group as many were suspected to have been exempted because their pretest score was felt to be not indicative of their true performance.

CONCEPTUALIZATION OF MODEL C

Whether or not the RI scores are elevated is the first question to be answered. We can begin to conceptualize the model by looking at Figure 3. All those students who have a pretest score below 40 are placed in the RI program as well as the regular curriculum, while all those who have a score of 40 and above are not allowed in the extra RI course and, hence, only receive the regular curriculum. After eight months of instruction, the posttest scores are obtained. The straight line of best fit is calculated for the comparison group. This line indicates the expected posttest performance for students at each pretest score. (See Figure 4.) If the line fits well, (correlation above .4) then we can proceed and assume that the straight line can be extended down into the range of scores of the treatment group which received RI. (See Figure 5.) We know, though, that the students below the cutoff not only received the regular curriculum but also received the RI curriculum. Therefore, the posttest scores of those receiving RI should be higher than if they would not have received RI. (See Figure 6.)

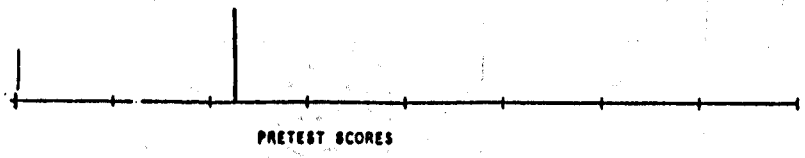


Figure 3. Selection of students into program, based on pretest score.

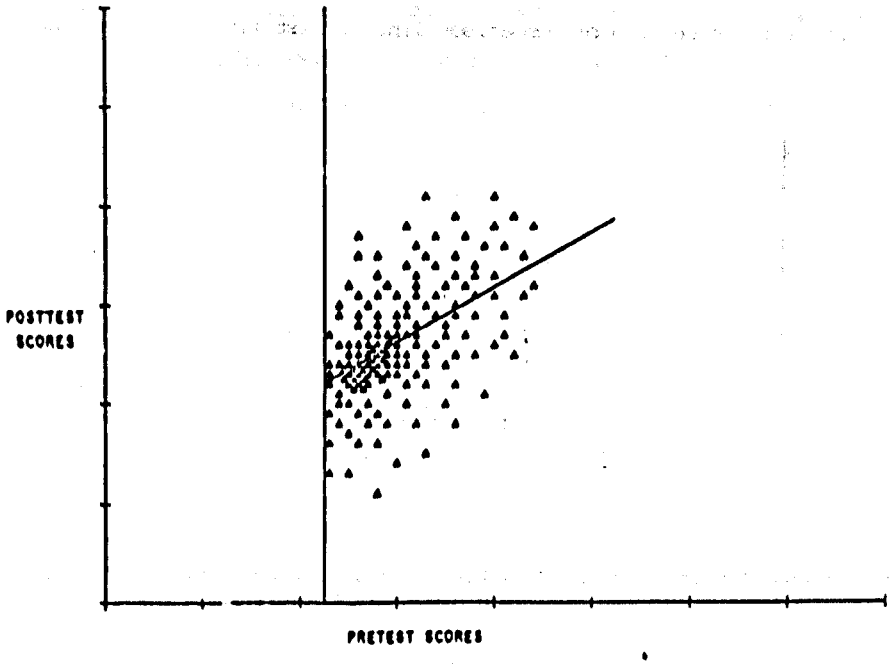


Figure 4. Line of best fit in comparison group.

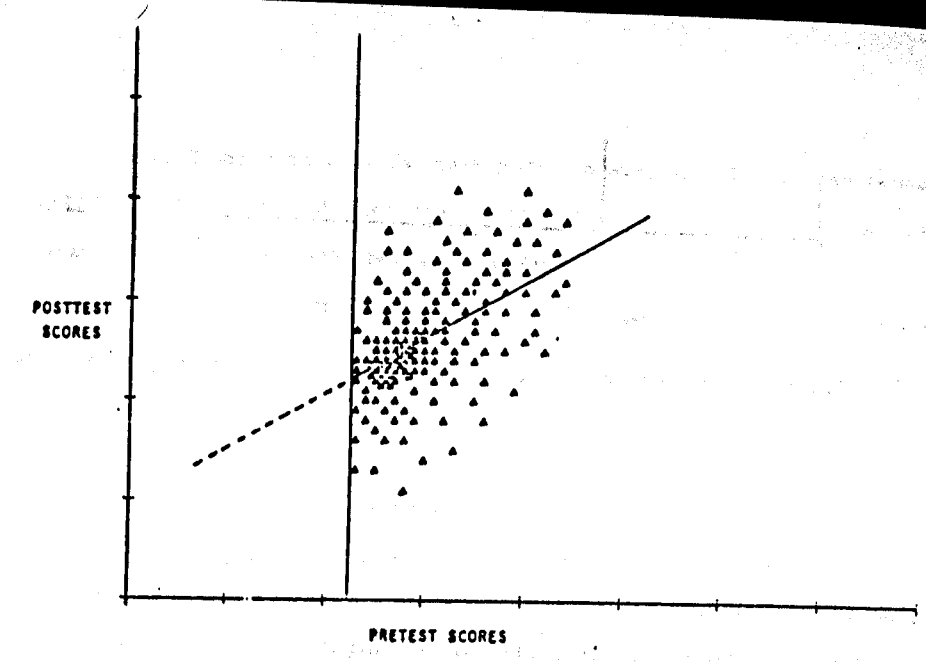


Figure 5. Extension of comparison line of best fit into treatment group.

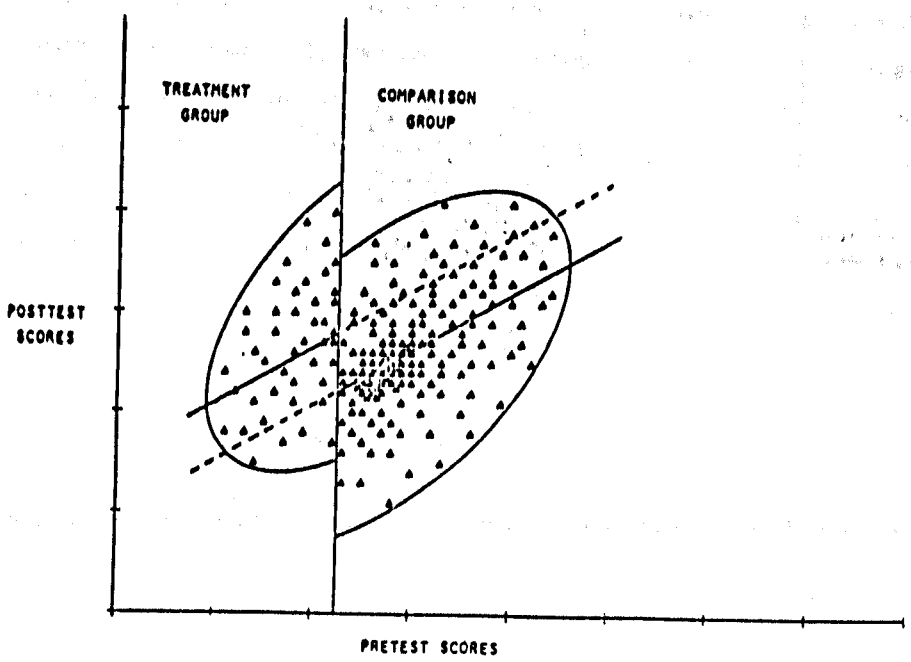


Figure 6. Model C illustration of treatment effect.

A second question of interest is whether the elevated effect was consistent across pretest scores. It might be that the RI program is especially effective in producing higher than expected gains for the lowest achieving students. (See Figure 7.) Or, the RI program may be especially effective for the highest students in the treatment. (See Figure 8.) Different program recommendations would, of course, result from these two different findings (findings which, by the way, would not surface in a Model A or Model B analysis). Thus, the second question of interest is, "Is the RI treatment differentially effective over the various pretest levels?" Another way to verbalize this interaction question is, "Is the RI line of best fit parallel to (exhibit the same slope as) the line of best fit for the comparison group?"

Model C, as any statistical question, can be tested with the general linear model. The full model contains all the information identified in the question (research hypothesis). Restrictions (identified in the question) are made on the full model, resulting in the restricted model. the difference in the number of pieces of information in the full and restricted models is equal to the number of restrictions. The general F-test formula is:

$$F = \frac{(R^2_{FULL} - R^2_{REST}) / (pieces_{FULL} - pieces_{REST})}{(1 - R^2_{FULL}) / (N - pieces_{FULL})}$$

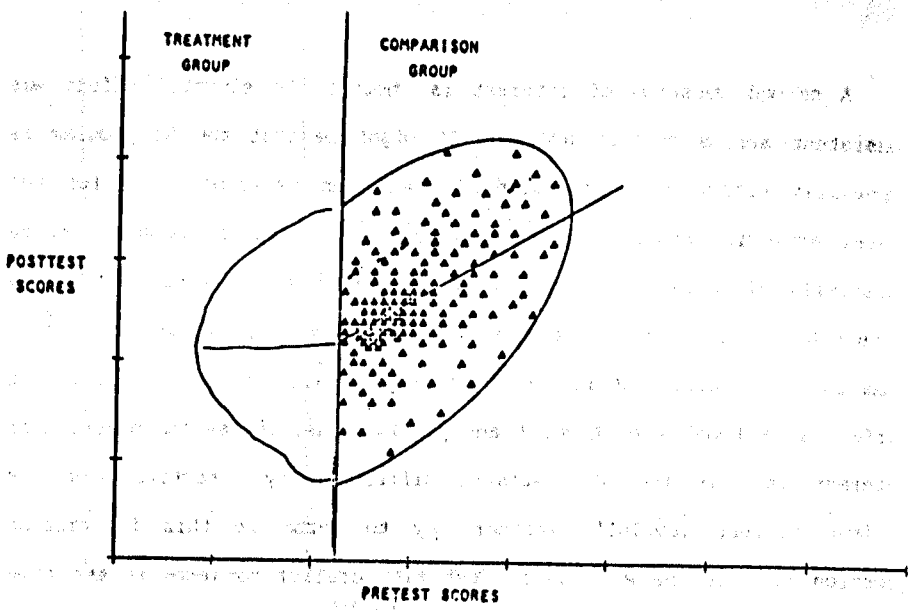


Figure 7. Model C illustration of treatment especially effective for l achieving treatment students.

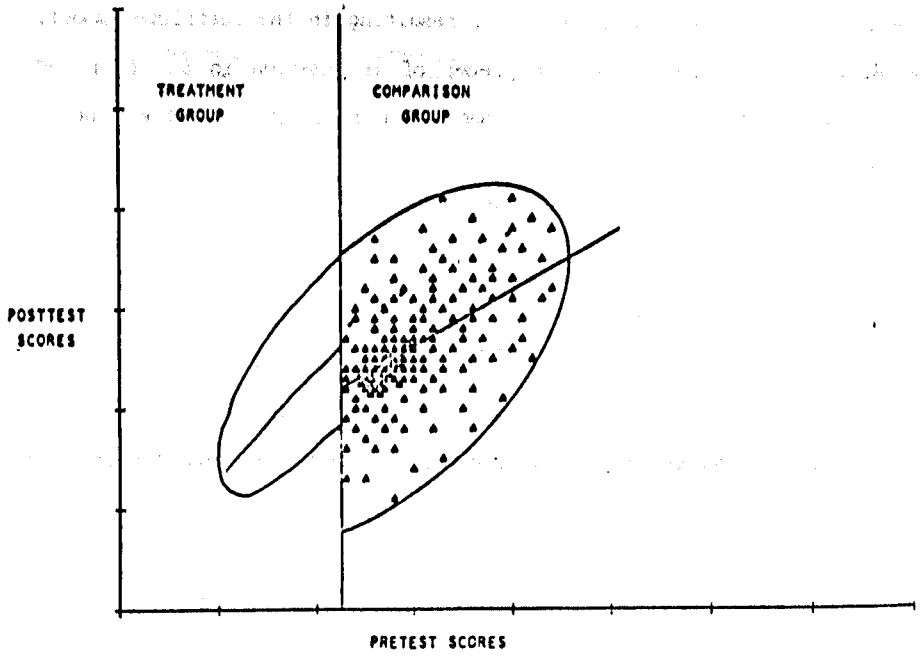


Figure 8. Model C illustration of treatment especially effective f high achieving treatment students.

MISCELLANEOUS DATA ANALYSIS TOPICS

Scales

All test information was transferred from percentiles to NCEs. NCEs are Normal Curve Equivalences which are a normal distribution transformation of percentiles. NCEs are an equal interval scale, therefore amenable to statistical manipulation. They have a mean of 50 and a standard deviation of 21.06.

Comparison groups

The comparison group should be receiving the regular curriculum received by the treatment group. In Dallas, most of the students above the 80th percentile enroll in an honors English course. Therefore, students above the 80th percentile were excluded from the analyses. Some students who should have been in the RI program because they had a qualifying pretest score below 40 were not given the special treatment. Before these students were combined with the regular comparison group, they were analyzed to see if they functioned differently.

Outliers

Students whose posttest scores were more than two standard errors of estimate beyond their predicted posttest score were eliminated from the analyses. The statistics used for a given student came from that student's group, RI comparison above 40, or comparison below 40.

MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 15, NUMBER 1, SUMMER 1986

Using Multiple Regression with Dichotomous Dependent Variables

Jerome D. Thayer
Andrews University

Introduction

Dichotomous variables are frequently encountered in multiple regression analysis, both as independent and dependent variables. A dichotomous independent variable is used to determine whether group membership is related to or will predict a certain outcome (i.e., whether gender predicts gpa). A dichotomous dependent variable is used to determine a combination of variables that will predict group membership (i.e., to predict dropping out of college).

Historically, whenever a dichotomous variable was studied as an independent variable with one dependent variable, a t-test, analysis of variance or analysis of covariance was conducted. When a dichotomous variable was studied as a dependent variable, discriminant analysis was used.

As multiple regression became more common, its advocates suggested that it could or should replace the t-test, ANOVA, ANCOVA or discriminant analysis in dealing with dichotomous variables by using coded variables.

Recently, however, Cox (1970), Goodman, (1978), Aldrich and Nelson (1984), and others have questioned the practice of using multiple regression when a dichotomous variable is used as the dependent variable. The most frequently suggested replacement for multiple regression is logistic regression.

In the introduction to Aldrich and Nelson (1984), it is suggested that ordinary regression analysis is not an appropriate strategy to analyze qualitative dependent variables, including those that are dichotomous. They go on to express the limitations of multiple regression very strongly:

> Perhaps because of its widespread popularity, regression may be one of the most abused statistical techniques in the social sciences. While estimates derived from regression analysis may be robust against errors in some assumptions, other assumptions are crucial, and their failure will lead to quite unreasonable estimates. Such is the case when the dependent variable is a qualitative measure rather than a continuous, interval measure. . . . For example we shall show that regression estimates with a qualitative dependent variable may seriously misestimate the magnitude of the effects of independent variables, [and] that all of the standard statistical inferences such as hypothesis tests . . . are unjustified (p. 9, 10).

The authors suggest that the failure of regression is "particularly troubling in the behavioral sciences" (p. 10), giving examples of qualitative dichotomous variables from the fields of political science, economics and sociology. Similar criticisms concerning dichotomous dependent variables are given strong emphasis in multiple regression textbooks aimed at economics and sociology, but popular regression textbooks in the behavioral sciences related to psychology and education do not express this same concern. For example, neither Cohen & Cohen (1975) nor Pedhazur (1982) deal with weighted least squares or logistic regression, two methods mentioned by multiple regression critics as preferable with dichotomous dependent variables. Both texts state that multiple regression can be used for and is mathematically equivalent to discriminant analysis when the dependent variable is a dichotomy (Cohen & Cohen, p. 442; Pedhazur, p. 687), but neither gives an indication that there are criticisms of this use. Tatsuoka (1971) states that in the dichotomous dependent variable case, multiple regression, discriminant analysis and canonical correlation are all mathematically equivalent and again, no indication is given of any criticisms of this approach.

Neter et al., (1983) list three problems that arise when the dependent variable is dichotomous: 1) non-normal error terms, 2) non-constant error variance, and 3) constraints on the response function. They state that even with binary dependent variables, ordinary least squares still provides

unbiased estimators under quite general conditions, and "when the sample size is large, inferences concerning the regression coefficients and mean responses can be made in the same fashion as when the error terms are assumed to be normally distributed" (p. 357). They add, however, that these estimators will not be efficient, giving larger variances than could be obtained with weighted procedures.

The solutions proposed to these problems include using weighted least squares to give constant error variance and using a transformation (such as logistic) that limits the response function to a range of 0 to 1.

In comparing the use of logistic regression or discriminant analysis with dichotomous dependent variables, Press and Wilson (1978) suggest that logistic regression is preferred except when the populations are normal with identical covariance matrices. They extend the criticisms of others to include situations in which dichotomous variables are used as independent variables. They state that logistic regression is valid for a wide variety of underlying assumptions including 1) all explanatory variables are multivariate normally distributed with equal covariance matrices, 2) all explanatory variables are independent and dichotomous, and 3) some are multivariate normal and some dichotomous whereas discriminant analysis is only valid under the first set of assumptions. These comments are not directed at multiple regression, but would apply in those situations where it is mathematically equivalent to discriminant analysis. Their conclusion is that logistic regression with maximum liklihood estimation is preferred to linear discriminant analysis. They state, however, that it is unlikely that the two methods will give markedly different results or yield substantially different linear functions unless there is a large proportion of observations whose x-values lie in regions of the factor space with linear logistic response probabilities near zero or one. They go on to say that logistic regression is preferred when the normality assumptions are violated, especially when many of the independent variables are qualitative.

The critics state that in addition to the predictions made by the regression equation with a dichotomous dependent variable, statistical tests are also invalid. This would include the F test of the overall model and the t values for each predictor in the model.

Cox (1970), in referring to the use of multiple regression with dichotomous dependent variables, states that "the use of a model, the nature of whose limitations can be foreseen, is not wise, except for very limited purposes" (p. 18). If these critics are correct, it appears as if researchers in education and psychology should discontinue the use of multiple regression in these situations.

Problem

This paper is an attempt to assess the meaning of the charges made against multiple regression and to suggest what the regression community in education and psychology can do to come to terms with critics of multiple regression. The purpose of this paper is not to evaluate the validity of the criticisms but to deal with some logical extensions of them. If these criticisms are valid, are t-tests, analysis of variance, analysis of covariance, discriminant analysis, canonical correlation, and any use of dummy variables in multiple regression also called into question?

The questions raised by this paper, then, are:

1. To what extent do these criticisms affect the validity of other comparable statistical procedures?
2. If other statistical procedures using different assumptions give identical results to multiple regression using dichotomous dependent variables, does this imply suspicion concerning the other procedures or suspicion concerning the validity of the criticisms or both?

Procedures and Findings

To examine the validity and/or seriousness of these criticisms, implications of this situation are considered by examining a set of data taken

from the A3 data set in Gunst & Mason (1980). This data set has 13 yearly observations with 14 variables. The year variable was dichotomized by letting the first 7 years be in one group and the last 6 years be the other group. The data is analyzed in 5 different cases with different arrangements of the dichotomous variable with one or two quantitative variables from this data set. The dichotomous variable is considered as both a dependent variable and an independent variable.

In Table 1 different combinations of quantitative and dichotomous independent and dependent variables where multiple regression has been used are presented with a listing of conventional alternative statistical methods and methods recommended by multiple regression critics. The critics suggest that in cases where a dichotomous dependent variable is used (cases 1 and 3) multiple regression is inappropriate. The approach taken in this paper is to compare the results of multiple regression in these cases with results of cases where multiple regression has not been attacked (cases 2 and 4).

Table 1

Possible Statistical Procedures to use with Different Combinations of Dichotomous and Quantitative Variables

Case	Dependent Variable	Independent Variable	Possible procedures
One Predictor			
1.	1 Dichotomous	1 Quantitative	Logistic regression Pearson correlation Pt. bis. correlation
2.	1 Quantitative	1 Dichotomous	t test Pearson correlation Pt. bis. correlation
Two+ Predictors			
3.	1 Dichotomous	2+ Quantitative/0+ Dichotomous	Logistic regression Discriminant analysis Multiple regression
4.	1 Quantitative	1+ Quantitative/1+ Dichotomous	Analysis of Covariance Multiple regression

Table 2 presents the results of the one predictor cases with the dichotomous variable as a dependent variable (case 1) and as an independent variable (case 2). In these situations the t value is the same whether the dichotomous variable is the independent or dependent variable. A one predictor model is the simplest case of multiple regression and the test of significance of the relationship is mathematically identical to an independent means t-test and a one-way ANOVA with two groups and the regression test of significance (t value) is the same whether the dichotomous variable is the independent or dependent variable. If a test of significance with a dichotomous dependent variable is invalid, then all tests of significance for an independent means t-test, a two-group one-way ANOVA and correlation/regression with an independent dichotomous variable are also invalid.

Table 2

One Predictor Examples

CASE 1: Multiple regression claimed to be invalid

Dependent variable = 2 (Dichotomous)
Independent variable = 3 (Quantitative)

$t_3 = -6.910$ -- same as case 2

CASE 2: Multiple regression is valid

Dependent variable = 3 (Quantitative)
Independent variable = 2 (Dichotomous)

$t_2 = -6.910$ -- same as case 1

Table 3 presents the results of the two predictor cases with the dichotomous variable as a dependent variable (case 3) and as an independent variable (cases 4a and 4b). Case 3 is a situation where multiple regression and discriminant analysis are both frequently used but is considered to be invalid by the critics of ordinary least squares due to the presence of a dichotomous dependent variable. The t values in case 3 are testing the significance of the relationship of each quantitative predictor with the

Table 3

Two Predictor Examples

CASE 3: Multiple regression claimed to be invalid

Dependent Variable = 2 (Dichotomous)
Independent Variables = 4 (Quantitative)
= 3 (Quantitative)

t_4 = -0.124 -- same as case 4a
t_3 = -6.480 -- same as case 4b

CASE 4: Multiple regression is valid

a. Dependent Variable = 4 (Quantitative)
Independent Variables = 2 (Dichotomous)
= 3 (Quantitative)

t_2 = -0.124 -- same as case 3
t_3 = -0.397

b. Dependent Variable = 3 (Quantitative)
Independent Variables = 4 (Quantitative)
= 2 (Dichotomous)

t_4 = -0.397
t_2 = -6.480 -- same as case 3

dichotomous dependent variable controlled for the other quantitative predictor. Cases 4a and 4b give identical t values to those found in case 3 for the relationship between the dichotomous variable (which is now one of the

independent variables and in a legitimate place according to assumptions of multiple regression) and the dependent quantitative variable. If the tests for which the t values in Case 3 are invalid, then the tests for which the t values in cases 4a and 4b are used are also invalid. The t values in cases 4a and 4b are the same as the square root of the F values that would be computed with a one-way analysis of covariance in which the independent quantitative variable was treated as the covariate and the independent dichotomous variable as the grouping variable. So therefore if Case 3 is invalid, then all one-way ANCOVA designs and any use of dummy variables in multiple regression would be invalid also.

Conclusion and Recommendations

It is clear from the above examples that the tests of significance are identical whether the dichotomous variable is an independent variable or a dependent variable. It appears, therefore, that if the critics of using multiple regression with a dichotomous dependent variable are to be taken seriously, they must also deal with all significance testing with t tests, analysis of variance, analysis of covariance, discriminant analysis, and any use of dummy variables in multiple regression. There may be other statistics reported in a multiple regression analysis, such as the standard error of estimate or predicted values for which the interpretations may not be appropriate when dichotomous dependent variables are used, but this paper will not deal with these issues.

BIBLIOGRAPHY

Aldrich, J. H. & Nelson, F. D. (1984). Linear Probability, Logit, and Probit Models. Beverly Hills, California: Sage Publications.

Cox, D. R. (1970). The Analysis of Binary Data. London: Methuen & Co.

Goodman, L. A. (1978). Analyzing Qualitative/Categorical Data. Lanham, Maryland: University Press of America.

Cohen, J. & Cohen, P. (1975). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Gunst, R. F. & Mason, R. L. (1980). Regression Analysis and its Application. New York: Marcel Dekker.

Neter, J. et al. (1983). Applied Linear Regression Models. Homewood, Illinois: Richard D. Irwin.

Pedhazur, E. J. (1982). Multiple Regression in Behavioral Research. New York: Holt, Rinehart, Winston.

Press, S. J. & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis, Journal of the American Statistical Association, 73, 699-705.

Tatsuoka, M. M. (1971). Multivariate Analysis: Techniques for Educational and Psychological Analysis. New York: John Wiley & Sons.

Relationship of Student Characteristics and Achievement in a Self-Paced CMI Application

Gerald J. Blumenfeld, Isadore Newman, Anne Johnson and Timothy Taylor
The University of Akron

Learner control of CBE applications has been an enticing topic of research. Reviews by Steinberg (1977) and Taylor (1976) indicate that effects upon achievement are equivocal when learner control has been compared with program or instructor control. The mixed results suggest the possibility of an interaction between certain aspects of instruction and characteristics of the learner, when the learner is permitted to control the program.

Current theory and data suggest that an important variable related to academic success is the student's perceived locus of control. Internal/external orientations have been shown to have a significant relationship to academic success (Coleman, et. al., 1966; deCharms, 1976). Behaviors exhibited by those having high internal or high external orientations (Crandall, et. al., 1965; Seeman, 1963; Seeman & Evans, 1962) appear to be closely related to successful use of opportunities that permit one to control the conditions of learning. It was hypothesized in this study that high internals would be more likely to explore and profit from learner control opportunities than would high externals. The I-E Scale developed by Rotter (1966) was considered to be an appropriate measure of this characteristic for college students.

A more direct measure of achievement-stricing behavior is the SSHA (Survey of Study Habits and Attitudes, Brown & Holtzman, 1967). This assesses the tendency of students to be prompt, to employ effective work methods, and to possess positive attitudes towards teachers and schooling. SSHA has been shown to be related to grade point average of college students (Brown & Holtzman, 1967; Desiderato & Koskinen, 1969) and to exam scores (Wen & Liu, 1976). It has also been shown that the SSHA and the I-E are related (Ramanaian et al., 1975).

It was hypothesized for the studies reported here that effective study habits would facilitate one's efforts to learn, and that this variable should interact with I-E when students are given an opportunity to excercise control. It was also hypothesized that these variables would be particularly salient in a self-paced CMI application where the instructor controlled the operating parameters during the second half of the course. Under such conditions, students who differ on these variables should exhibit even greater differences on achievement as the course progresses.

METHOD

Subjects

Subjects were students enrolled in a junior level college course on educational measurement. Study A was conducted during the spring quarter of 1978;

*This article is based upon a paper presented at the American Psychological Association

Study B was conducted during the spring semester of 1979. The University had changed from quarters to semesters and the course went from 3 quarter hours to 2 semester hours. Requirements of the course and sequencing of activities remained the same.

Measures listed below were secured for 102 of 133 students enrolled during 1978 and for 86 of 125 enrolled during 1979. Most of the students not included in the samples withdrew from the course very early in the term. A few students were absent on the days the I-E and SSHA were administered.

Instruments

Rotter's I-E Scale and the Brown-Holtzman SSHA were administered during regular class meetings. Students were given individual feedback about these measures at the end of the term. A brief description of each is presented below:

1. I-E Scale. This scale contains 29 forced-choice items, including six fille items, and was keyed so that a high score indicated a high internal orientation

2. SSHA. This inventory contains 100 items grouped into the following subscal

   a. Delay Avoidance (DA). Lack of procrastination.
   b. Work Methods (WM). Effective study procedures.
   c. Study Habits (SH). DA plus WM
   d. Teacher Approval (TA). Attitude towards teachers and their behavior.
   e. Education Acceptance (EA). Attitude towards educational practices.
   f. Study Attitudes (SA). TA plus EA
   g. Study Orientation (SO). SH plus SA (overall measure)

3. Comprel nsive Exams. Achievement in each of four units of work was measure by thirty item selection type exams. Two or three alternate forms were available for each unit. Exams used were a regular part of the course. Item analys indicate acceptable quality. Measures of reliability have ranged from .70 to more than .90. Method of estimating reliability, number of students involved, and term when analysis was conducted, varied from one set of unit exams to ano

4. GPA. Overall grade point average at end of spring quarter was obtained fr the registrar's records. This measure has been shown to be correlated with ac demic achievement in the measurement course (Blumenfeld, et al. 1975) and with research on learner control in CAI (Taylor, 1976).

Procedures

Student behavior and achievement was examined under conditions imposed by self-paced computer managed instruction applied to an undergraduate educationa measurement course. A brief description of the course and the computer progra is given below. More detailed accounts can be found in Blumenfeld, et al. (19 and Blumenfeld, et al. (1977).

1. Measurement Course. Emphasis is placed upon evaluating the effectiveness of instruction and upon developing coordinated sets of instructional objective instructional procedures, and measurement procedures. The course is divided

into four units and one teaching project. The teaching project is not self-paced and student behavior related to this aspect of the course was not included in the analysis. The units are divided into modules - three modules per unit. Students are given ten to fifteen behaviorally stated objectives for each module. They are required to take module study quizzes via computer terminals on each unit they study prior to taking comprehensive exams. Study quizzes are taken outside of regular class time and are scheduled by the student at his or her convenience. Comprehensive exams are given during regular class time during six predetermined sessions distributed throughout the term. Criterion for passing a unit exam is 80%. A student may take a second exam on each unit if he fails to pass the first time. A few target points awarded at the beginning of the term, to encourage students to get started, permit a few students to pass unit I with only 70% correct. Course grade is determined by the number of units the student passes. If the student passes four units, a grade of A is recorded; three units, a grade of B is recorded, etc. Minus grades are given if students achieve 70% but not 80%. The student can decide to work on all four units or to stop after one. Upon request, incompletes are awarded to permit a student to complete one additional unit. Only work completed during the spring term was included in the analysis.

The topics included in modules one thru six are repeated in modules seven thru twelve. Objectives in the first six modules include critical concepts and less difficult tasks. Objectives in the last six modules include more advanced ideas and more difficult tasks.

2. Computer Program. The program contains twelve quizzes with each quiz containing twelve items. A pool of five selection type items is included for each objective. When a student signs on, the program randomly orders the objectives and randomly selects one item for each objective. Emphasis is provided by including two five item pools for some objectives and repeating these objectives. After a correct answer, the student is so informed. Appropriate page references for three books follow both correct and incorrect answers. If an incorrect answer is given, the student is informed as to why the answer is not correct. Correct answers are not given, but the student is provided with some direction for reconsidering the problem. At the end of the quiz the student can see a list of objectives related to the items answered incorrectly.

Criterion for passing is ten correct answers. If the student meets the criterion the program advances the student to the next module. If the student fails to meet the criterion a second or third study quiz on that module is required. A delay of ten minutes per error is imposed before the student is permitted to take another quiz. Students failing a module quiz for the third time are advanced to the next module. A student who fails three quizzes on two consecutive modules is not permitted to continue until he obtains a "password" from the instructor. After the student has completed module six, control of the computer program is given to the student. The student decides which module to go to, how many times to take a quiz on that module, and in what order to repeat modules if he so chooses. The student can avoid any delay imposed earlier because of errors.

RESULTS

Intercorrelations of Measures

Tables 1, 2 and 3 are divided into parts A and B and correspond to 1978 and 1979 data, respectively. Table 1 lists the intercorrelations of the SSHA

TABLE I

Intercorrelation of the SSHA Scales, I-E and GPA

		DA	WM	SH	TA	EA	SA	SO	IE	GPA
STUDY A	DA		.59	.89	.50	.67	.63	.82	.07	.33
	WM			.89	.65	.63	.68	.84	-.08	.36
	SH				.65	.73	.73	.94	-.01	.39
	TA					.77	.94	.84	-.06	.23
	EA						.94	.89	-.12	.30
	SA							.92	-.10	.29
	SO								-.06	.35
	IE									.09

(r ≥ .195 significant at .05 level)
(r ≥ .254 significant at .01 level)

STUDY B	DA		.68	.92	.38	.60	.53	.81	.30	.21
	WM			.91	.57	.55	.62	.85	.26	.11
	SH				.52	.63	.63	.91	.30	.18
	TA					.67	.92	.77	.42	.13
	EA						.91	.83	.40	.09
	SA								.45	-.02
	SO								.43	.09
	IE									-.13

(r ≥ .212 significant at .05 level)
(r ≥ .277 significant at .01 level)

scales, I-E and GPA. It is interesting to note that in study A all correlations are significant except those involving I-E. In study B all correlations are significant except those involving GPA.

Table 2 indicates the relationship of the variables described above to unit exam scores. It can be observed that while GPA is significantly related to unit exam scores in both studies, I-E and SSHA do not possess that consistency. I-E is related to unit exams in study A but not in study B. SSHA is not related to unit exams in study A but many of the correlations approach significance in study B.

Regression Analysis

To test the original hypotheses that I-E orientation and study skills would be salient variables further analyses were conducted using SH since the relationship of the other scales to unit exams was not significant. Full and restricted regression models were used to examine the predictiveness of GPA, SH, I-E, and (SH * I-E) when the criterion was unit exam score. Regression models were computed for each of the four unit exams. GPA was included in all models. Therefore, tests conducted determined whether or not SH, I-E and (SH * I-E) could account for a significant amount of criterion variance above and beyond that accounted for by GPA. The interaction (SH * I-E) was found not to be statistically significant, nor was SH.

In study A I-E was found to be significant at the .01 level for units I, II, and III and at the .05 level for unit IV. However, I-E did not account for a significant amount of criterion variance beyond that accounted for by GPA in study B. The multiple R^2 for the full and restricted models are given in Table 3.

Ad Hoc Analysis

Trends

It was hypothesized that the effects of variation in locus of control and study habits upon student performance would increase as the term progressed. Therefore, intercorrelations across modules and units were examined to determine if any trends could be detected. In study A the correlation matrix indicated that DA was the most likely scale to generate a significant trend. Cumulative exam scores across the four units were recorded for both the first and fourth quartile groups on the delay avoidance scale. Traditional analysis of variance for trend was inappropriate because of extreme heterogeniety of variance. Therefore, log-log transformations were made for each student's cumulative exam score curve.

The slope of the regression line for each of these log-log transformations was computed. This was used as a measure of trend. The means of the slopes for the two groups were .86 and .65; the variances were .05 and .08. Students who scored high on DA had the higher mean slope. A test of these values indicated that the difference between the means of the slopes was significant at the .01 level. Obtained t was 2.776 with df = 48. This trend was not found to be present in the data obtained in study B.

Use of CMI Program

No directional hypotheses with respect to student utilization of the CMI program were formulated. However, it is reasonable to assume that successful

TABLE 2

Correlations of SSHA Scales, I-E and GPA with Unit Exam Scores

	Unit Exam	DA	WM	SH	TA	EA	SA	SO	IE	GPA
STUDY A	I	.03	.20	.13	.11	-.04	.04	.09	.29	.46
	II	.10	.15	.14	.05	.01	.02	.09	.35	.31
	III	.18	.12	.16	.11	.06	.10	.13	.33	.37
	IV	.26	.16	.23	.08	.11	.11	.19	.19	.38
		($r \geq .195$ significant at .05 level) ($r \geq .254$ significant at .01 level)								
STUDY B	I	.14	.21	.19	.05	.16	.12	.19	-.04	.33
	II	.16	.23	.21	-.02	.16	.07	.17	-.13	.48
	III	.26	.25	.28	.02	.22	.13	.24	-.07	.50
	IV	.19	.21	.22	-.02	.18	.08	.16	.04	.34
		($r \geq .212$ significant at .05 level) ($r \geq .277$ significant at .01 level)								

TABLE 3

Regression Analysis Models
Indicating Significance of I-E

Criterion: Unit Exam	R^2 of Full Model (GPA, SH, IE)	R^2 of Restricted Model (GPA, SH)	F	df	p
I	.28	.21	8.577	1,98	<.01
II	.21	.10	13.502	1,98	<.001
III	.23	.13	11.744	1,98	<.001
IV	.18	.15	3.158	1,98	<.05

Comparison of full and restricted models indicates that I-E accounts for a significant amount of exam score variance beyond that accounted for by GPA.

I	.129	.126	.278	1,83	<.60
II	.231	.222	.940	1,83	<.34
III	.242	.236	.574	1,83	<.45
IV	.158	.157	.038	1,83	<.85

Comparison of full and restricted models indicates that I-E does not account for a significant amount of exam score variance beyond that accounted for by GPA.

students will utilize learning resources differently than less successful students. The twenty five students who had the highest score on a combination of I-E and DA were identified along with the twenty five students who had the lowest score on this variable. The variable was obtained by multiplying each student's I-E score by his DA score. The mean number of quizzes per module and the mean number of minutes per module were computed for each of these groups. Only the first nine modules were considered because a very small percentage of students worked on Unit IV. This fact will be considered later. Nine out of nine times the high group's mean number of quizzes per module was greater than the mean of the low group. Eight out of nine times the mean of the high group's number of minutes per module was greater than the mean of the low group. On the average, members of the high group took more study quizzes, but spent less time per quiz than did members of the low group. High students were not only practicing more but also distributing the practice across a greater number of examples. Once again, it was found that this relationship did not occur in the data collected from study B.

Accurate records of when students took module quizzes and unit exams were obtained for study B. This data was examined several ways, but no consistent relationships between student characteristics and the utilization of the CMI program were observed.

Discussion

It is important to note at the beginning of this discussion that an unusually small percentage of students worked on unit IV during the termsstudy A was conducted. For example, thirty six percent of the students listed on our first day roster for the previous quarter worked throughout the term and received a grade of A or A- for the course. In study A only sixteen percent of the students listed on our first day roster worked throughout the term and received a grade of A or A-. In study B 33% earned a grade of A or A-. A non-scientific explanation is that the 1978 students suffered thru a very difficult winter. When the sun finally appeared during the spring quarter, students stopped working on all non-required school tasks. We observed this sudden cessation of study and were given this answer when we raised questions about it.

It was assumed in 1978 that the small number of students completing unit IV would tend to restrict the range of scores involved and not invalidate the results. The failure to replicate the results in 1979 leads one to other speculations. For example, the 1979 students had 50% more time to do the same amount of work and were not harrassed by bad weather and school closings. It is possible that differences in I-E and SSHA interact with conditions of stress and high demands. When such conditions are not present, as in study B, all students have time to do the job even if differences in ability and motivation exist.

This is an attractive hypothesis, but it is also suspect because of the change in the observed relationships between I-E and SSHA scales. Weather and length of term should not have had an affect here. It is also the case that the relationships between I-E and SSHA in study B are more consistent with the data reported by Ramanaiah (1979).

Study A supported the conclusion that the I-E and SSHA scales tapped important student characteristics when course structure permitted students to control pace, practice conditions, utilization of resources and total amount of material to be studied and mastered. Study B does not support those conclusions. Only additional replications will provide help in deciding which set of data should command one's confidence.

At least two things should be considered when looking at the results of study A and study B. One is that apparently the most relevant psycho-social variables have not been adaquately identified. The second, and more importantly is that the different results give further support for the necessity to replicate. The two studies reported were conducted by the same researcher, on very comparable students, in highly similar settings, yet produced divergent results. These varying results indicate the potential pitfall of generalizing results based on only one study.

When trying to identify the relevant learning characteristics in a natural setting, the potential interactions and the types of relationships between variables are enormous. What may be needed to map out many of these possible relationships, develop a matrix, and systematically develop studies to investigate the relationship between these variables and learning. One may take a particular model such as suggested by McGuire (1960) and Whiteside (1964) which takes the position that when one is trying to account for complex behavior, one has to look at atleast three classifications of behavior. One is the person variables which includes things such as personality, intelligence, sex roles, learning characteristics, etc. The second is the characteristics of what is to be learned. Suppes (1966) and Gagne' (1965) have given excellent examples of how to deliniate the components of what is to be learned through a task or job analysis. The third is the environmental or context variables. These would include such things as the structure as well as the environment of the learning situation, interactions with peers, expectations produced by the environment (significant others within the environment). This three dimensional matrix may facilitate the identification and systematic investigation of the variables which may influence and/or "cause" the differential effectiveness of "learning" as reported in the literature.

---

*However, one must be very careful of over generalizing to other samples before independent replications are conducted. The authors have collected replication data which they expect to present at a future time in conjunction with the findings of this paper.

References

Brown, W. F., & Holtzman, W.H. Survey of Study Habits and Attitudes: Manual. New York: The Psychological Corporation, 1966.

Blumenfeld, G., Hirschbuhl, J., & Taylor, T. "CMI applied to a self-paced undergraduate tests and measurements Course: First Year Report," Summer Meeting of the Association for the Development of Computer Based Instructional Systems, 1975.

Blumenfeld, G., Hirschbuhl, J., & Taylor, T. Computer managed measurement course. In S. K. Siders (Ed.), Restructuring Teacher Education: Resources for Curriculum Planning. Ohio Department of Education, 1977.

Coleman, J.S., Campbell, E. Q., Hobson, C.J., McPortland, J., & Mood, A.M. Equality of educational opportunity, Washington, D.C.: U.S. Department of Health, Education and Welfare, 1966.

Crandall, V.C., Katkovsky, W., & Crandall, V.J.. Children's beliefs in their own control of reinforcements in intellectual-academic achievement situations. Child Development, 1965, 36. 91-109.

deCharms, R. Enhancing motivation: Change in the Classroom. New York: Irvington Publishers, Inc., 1976.

Desiderato, O. and Kiskinen, P. Anxiety, study habits, and academic achievement. Journal of Counseling Psychology, 1969, 16, 162-165.

McGuire, C. "The Textown Study of Adolescence," Texas Journal of Science, 8, 1956.

Ramanaiah, N.V., Ribich, F.D., & Schmeck, R.R. Internal external control of reinforcement as a determinant of study habits and academic attitudes. Journal of Research in Personality, 1975, 9, 375-384.

Rotter, J.B. Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs, 1966, 80 (1, whole No. 609).

Seeman, M., & Evans, J.M. Alienation and learning in a hospital setting. American Sociological Review. 1962, 27, 772-782.

Seeman, M. Alienation and social learning in a reformatory. American Journal of Sociology, 1963, 69, 270-284.

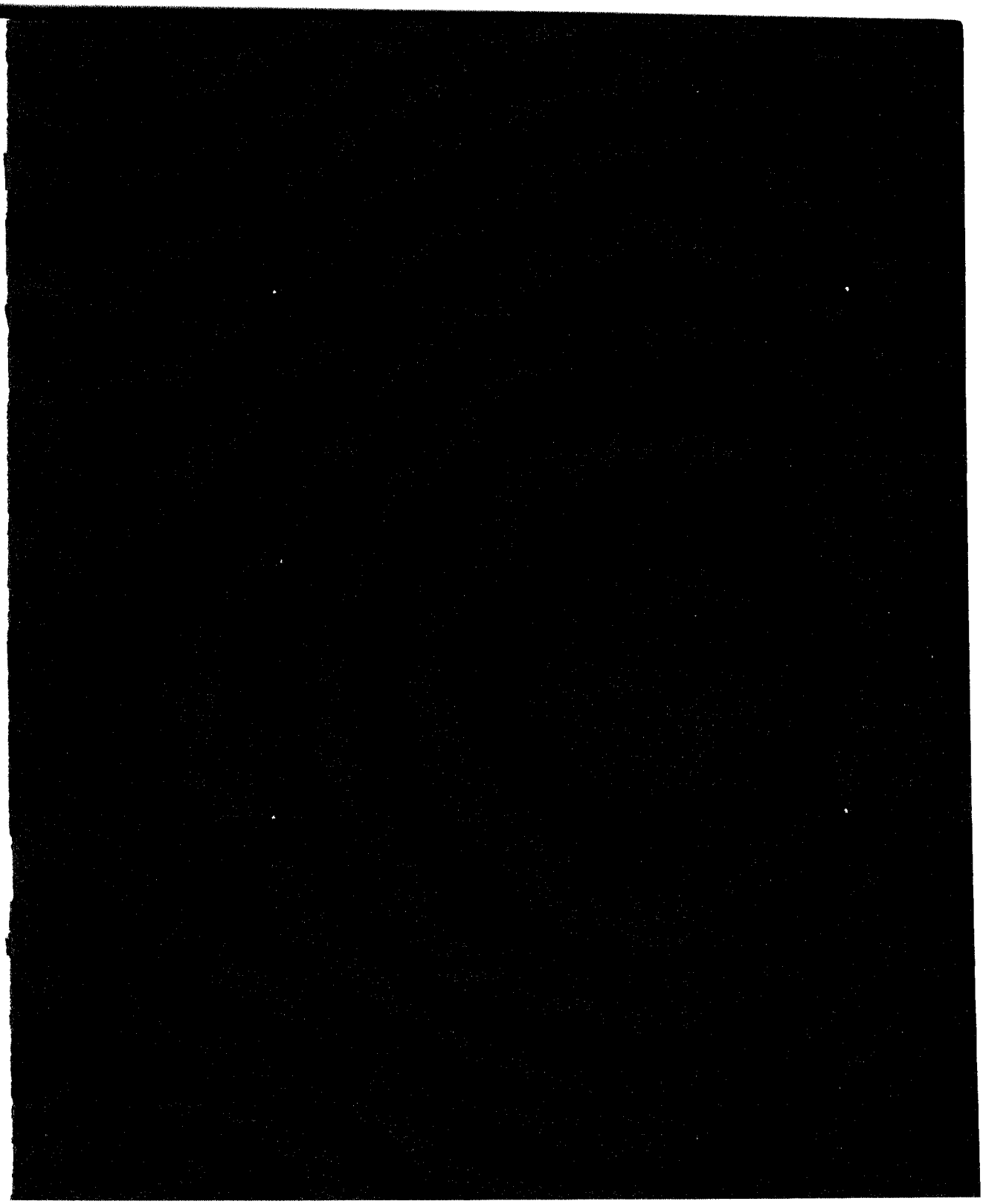
Steinberg, E.R. Review of student control in computer-assisted instruction. Journal of Computer-Based Instruction, 1977, 3, 84-89.

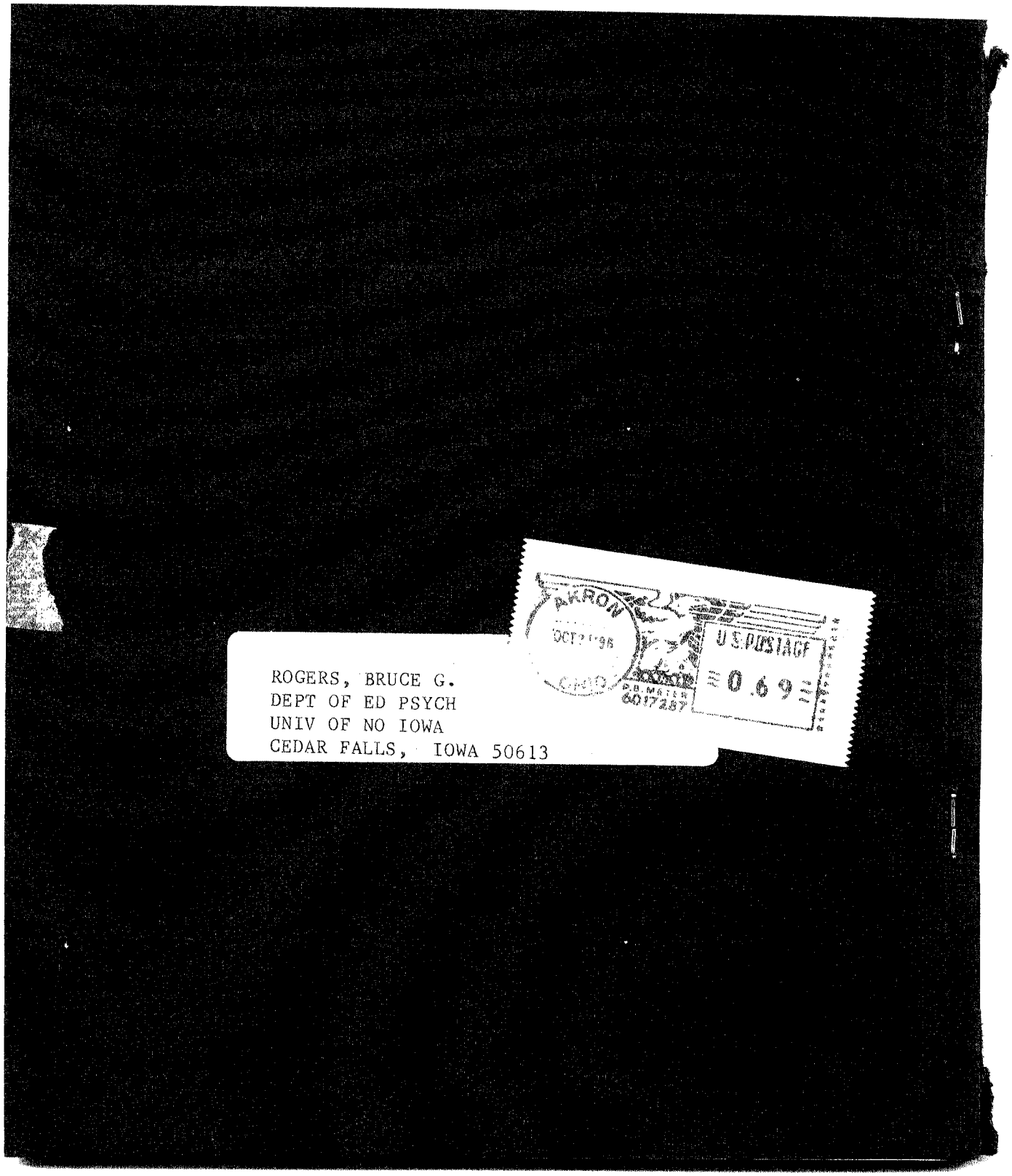
Suppes, P., L. Hyman and M. Jerman "Linear Structural Models for Reponse and Latency Performance in Arithmetic" Technical Report 100, Standford, CA Inst. for Math. studies in the social sciences 1966.

Taylor, R. The effect of behavioral objectives upon selected aspects of learner control in a computer-based instructional setting. Unpublished doctoral dissertation, The University of Akron, 1976.

Wen, S. & Liu, A. The validity of each of the four scales of the survey of study habits and attitudes for each of two samples of college students and under each of two treatment conditions involving use of released class time. Educational and Psychological Measurement, 1976, 36, 565-568.

Whiteside, R. Dimensions of teacher evaluation of academic achievement. Unpublished doctoral dissertation. The University of Texas, Austin, Texas 1964.




AKRON
OHIO
U.S. POSTAGE
0.69
ROGERS, BRUCE G.
DEPT OF ED PSYCH
UNIV OF NO IOWA
CEDAR FALLS, IOWA 50613