On the Investigation of Alternative Regressions by Principal Component Analysis

By DOUGLAS M. HAWKINS

University of the Witwatersrand, Johannesburg, South Africa

SUMMARY

In a multiple regression problem, let the $p \times 1$ vector $\mathbf{x}$ consist of the dependent variable and $p-1$ predictor variables. The correlation matrix of $\mathbf{x}$ is reduced to principal components. The components corresponding to low eigenvalues may be useful in suggesting possible alternative subregressions. This possibility is analysed, and formulae derived for the derivation of subregressions from the principal components.

Keywords: MULTIPLE REGRESSION; SUBREGRESSION; PRINCIPAL COMPONENT ANALYSIS

1. INTRODUCTION

IN multiple regression analysis we have at our disposal a number of predictor variables from which we wish to predict the dependent variable. Several situations may arise:

(1) all predictors may be essential for accurate prediction of the dependent variable;
(2) some predictors may have no predictive value, and so may be dropped from the problem altogether;
(3) there may exist subsets of the predictors, either partially or totally disjoint, which provide a prediction essentially as accurate as the full set of predictors.

An essential feature of this third situation is the possible existence of alternative subsets of predictors having essentially the same multiple correlation with the dependent variable. For example if there is an exact multicollinearity amongst some of the predictors, then any one of these predictors may be eliminated from the problem. From the mathematical point of view it is immaterial which of the possible predictors is eliminated. The user, however, may have some feelings about which predictors he would prefer to use.

Thus where several alternative subsets of predictors exist all of which provide good predictions of the dependent variable, the user should ideally be made aware of all these subsets. He can then decide which subset to use in actual practice on the basis of nonstatistical criteria such as ease or cost of measurement of the predictors.

There are a number of techniques for investigating the possible use of a subset of the prediction.

Amongst these are:

(1) Stepwise regression and its relatives—forward selection and backward elimination (Efroymson, 1965).

In stepwise regression, we start from subregression using only a subset of the predictors. If any predictor in the subset contributes insignificantly to the regression (as measured by Anova) it is removed from the subset. If any predictor not in the subset would contribute significantly, it is introduced. The procedure then repeats. In this way, starting from any subset, the process converges to some subset that is locally stable. Stepwise regression is widely used, but has certain defects:

(a) The final subset will, in general, depend on both the initial subset and the significance levels used in the Anovas. In some problems stepwise regression may terminate far from the global optimum.

(b) Where alternative good subsets of predictors exist, the procedure will select one arbitrarily. The user thus remains unaware of the existence of the other subsets, and so can exercise control over the choice of subset only with some difficulty.

(2) "Best subset" (Garside, 1971; Beale *et al.*, 1967). Here, all possible subsets of the predictors are considered. This is the only technique guaranteed to find all good subsets of predictors. The number of possible subregressions may be large. If there are k predictors, then there are 2^k possible regressions, and if k is large, then the amount of analysis required to locate all subsets will be excessive.

(3) Factor analysis. Here the predictors are subjected to a factor analysis. Several factors emerge and the remainder of the variation is ascribed to near multicollinearities which may be ignored. The factors are rotated to obtain a simple structure in which each factor identifies with some minimal subset of the predictors. Regression of the dependent variable on the factors may then suggest suitable subsets of the predictors. This technique is used in Massy (1965), Daling and Tamura (1970) and a number of other practical studies.

Two objections may be raised to this technique:

(a) There is no guarantee that the dependent variable is dependent on the factors rather than on the near multicollinearities which have been ignored. Examples in which this is the case are given in Browne (1969) and Hotelling (1957).

(b) The technique suggests one or more possible subsets of predictors, but gives no explicit information on the number or composition of the alternative good subsets.

Thus the techniques of stepwise regression and factor analysis are defective in that the existence of alternative good subsets of predictors is not easily inferred. There seem to be a need for some technique that may be used as an auxiliary to existing methods which will expose the underlying structure of the relationships amongst the predictors and the dependent variable. It should enable alternative good subsets to be found easily, and should indicate multicollinearities that imply eliminable predictors.

In this paper we discuss a technique that is directed at these problems. It is primarily a descriptive aid for identifying likely subsets, and is useful for establishing the amount of leeway the user has for forming subregressions. In the final analysis, a number of subsets will be examined. The number will, however, generally be considerably lower than would be required by the best subset method.

2. Notation and Results

Suppose the $p \times 1$ vector $\mathbf{x} = (x_1 x_2 \dots x_p)'$ is made up of the dependent variable x_1 and $p-1$ predictors $x_2, \dots, x_p$, and assume the x_i are scaled to zero mean and unit variance in the given data.

Let $\mathbf{R} = (r_{ij})$ be the $p \times p$ correlation matrix of $\mathbf{x}$, which we assume to be nonsingular. The problem of multiple regression consists of finding a hyperplane:

$$x_1 - b_2 x_2 - b_3 x_3 \dots - b_p x_p = 0 \qquad (1)$$

such that the sum of squared deviations along the x_1 axis (s.s.d.) is minimized. A subregression not using all the predictors is also of the form (1), where the s.s.d. is to be minimized subject to the constraint that certain of the b are zero. This measure of goodness of fit, the mean squared deviation along the x_1 axis, will be termed the x_1 norm in this paper.

Let us now alter our criterion of goodness of fit. Suppose that the distance from the vector $\mathbf{x}$ to the hyperplane (1) is measured, not along the x_1 axis, but along the norm to the hyperplane.

This vertical distance is

$$\frac{x_1 - b_2 x_2 - \ldots - b_p x_p}{(1 + b_2^2 + \ldots + b_p^2)^{\frac{1}{2}}} \tag{2}$$

or

$$a_1 x_1 + a_2 x_2 + \ldots + a_p x_p,$$

where

$$l^2 = 1 + b_1^2 + \ldots + b_p^2, \quad a_1 = 1/l \quad a_i = -b_i/1 \quad (i \geqslant 2),$$

where l is the length of the coefficient vector

$$(1, -b_2, -b_3 \ldots - b_p) \quad \text{and} \quad a_1 = \cos\theta,$$

where θ is the angle between the x_1 axis and the normal to the hyperplane. The mean squared deviation along the normal to the hyperplane will be termed the vertical norm of the hyperplane.

If s^2 is the x_1 norm of any hyperplane and λ the vertical norm, then clearly

$$\lambda = a_1^2 s^2 < s^2$$

and

$$s^2 = l^2 \lambda. \tag{3}$$

Intuitively we might feel that since s^2 and λ both measure the fit of the hyperplane, a small value of λ should imply a small value of s^2 and vice versa. This is not necessarily so, as we can see from (3). Low s^2 implies low λ, since $0 < a_1^2 < 1$. However given λ, s^2 may be arbitrarily large if a_1^2 is arbitrarily small. (We note that a low λ with a low a_1^2 corresponds to a near multicollinearity amongst the predictors.) Thus a linear combination of the x_i which has a low λ identifies a low s^2 predictor of x_1 if a_1^2 is large, or a near multicollinearity amongst the predictors if a_1^2 is small.

The situation in which there are alternative subregressions is characterized by the fact that there is a space of hyperplanes all of which lie "close" to the data in the x_1 norm. There are infinitely many hyperplanes in this space, and if a predictor is eliminable, then there will be one or more hyperplanes in this space which run parallel to the axis of that predictor. The space is characterized by some basis of hyperplanes. From the basis, other hyperplanes can be deduced. Thus we can shed some light on the problem of alternative subsets if we can find informative, easily interpreted basis hyperplanes.

The problem of finding such a basis in the x_1 norm is not solved. The problem of finding a basis in the vertical norm, however, is well known. It proceeds as follows:

Find $\mathbf{a} = (a_1, a_2, \ldots, a_p)$ such that $\mathbf{aa}' = 1$ and $\mathbf{ax}$ has minimum variance. The vector $\mathbf{a}$ satisfying this minimum is the eigenvector corresponding to the smallest eigenvalue of $\mathbf{R}$.

The vertical norm λ is the corresponding eigenvalue. Having found this vector, we may now proceed to find another vector, orthogonal to the first which minimizes the vertical norm.

Extending this process, suppose we reduce $\mathbf{R}$ to canonical form. Let

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$$

be the eigenvalues, and let the ith eigenvector, corresponding to λ_i, be

$$\mathbf{a}_i = (a_{i1}, a_{i2} \ldots, a_{ip}).$$

Consider the hyperplane defined by $\Sigma c_i \mathbf{a}_i \mathbf{x} = 0$ with $\Sigma c_i^2 = 1$. The vertical norm will be $\lambda = \Sigma c_i^2 \lambda_i$, and the x_1 norm $s^2 = \lambda/\{\Sigma c_i a_{i1}\}^2 > \lambda$. Now if s^2 is small, then so also is λ. But λ is a weighted average of the λ_i, and if it is small, then the weights on the large λ_i must be low. Thus if s^2 is small, then the hyperplane loads predominantly on the eigenvectors corresponding to low eigenvalues.

This implies that if any hyperplane has low s^2, then it will be closely approximated by some linear combination of the eigenvectors corresponding to low eigenvalues. Thus we infer that if $\mathbf{R}$ has say m eigenvalues which for the purpose at hand are small, then the space of hyperplanes having low s^2 values will have a dimensionality not greater than m. Furthermore, the space spanned by the eigenvectors contains close approximations to any hyperplane having low s^2. This means that the reduction of $\mathbf{R}$ to canonical form may be used descriptively to acquire some feeling for the structure of interrelationships amongst the x_i.

3. Derivation of Regressions

We now consider the use of the eigenvectors to compute regression and sub-regression coefficients. Define the $p \times p$ matrix $\mathbf{D}$ by

$$d_{ij} = a_{ij}/\sqrt{\lambda_i},$$

and let

$$\mathbf{y} = (y_1 y_2 \ldots y_p)' = \mathbf{Dx}.$$

Note that $d_{i1}^2 = a_{i1}^2/\lambda_i$. Thus by (3), the x_1 norm of the ith eigenvector is $1/d_{i1}^2$.

The y_i are mutually uncorrelated, and have zero mean and unit variance. Regarded as functions of x_i the y_i form an orthogonal basis for the space R^p of $\mathbf{x}$ vectors. Consider the equation

$$\sum_{i=1}^{p} c_i y_i = 0,$$

with

$$\sum_{i=1}^{p} c_i d_{i1} = 1. \tag{4}$$

Written in terms of the x_i, this is

$$\sum_{j=1}^{p} \left\{ \sum_{i=1}^{p} c_i d_{ij} \right\} x_j = 0, \tag{5}$$

the condition $\Sigma c_i d_{i1} = 1$ implying that the coefficient of x_1 is 1.

Thus (5) may be written

$$x_1 = -\sum_{j=2}^{p} x_j \left\{ \sum_{i=1}^{p} c_i d_{ij} \right\}$$

and viewed as a predictor of x_1, (5) will have residual variance

$$\operatorname{var} \Sigma c_i y_i = \sum_{i=1}^{p} c_i^2.$$

Suppose we wish to find the conventional multiple regression using some subset of the predictors—say

$$x_{k+1}, x_{k+2}, \ldots, x_p.$$

This problem can be formulated as follows: minimize $\sum_i c_i^2$, subject to the constraints

$$\sum_i c_i d_{i1} = 1$$

and

$$\sum_i c_i d_{il} = 0, \quad l = 2, 3, \ldots, k. \tag{6}$$

Let $\mu_1, \mu_2, \ldots, \mu_k$ be k Lagrange multipliers. Then our objective is to minimize

$$S = \Sigma c_i^2 - 2\mu_1 \left\{ \sum_i c_i d_{i1} - 1 \right\} - 2 \sum_{l=2}^{k} \mu_l \sum_i c_i d_{il}$$

$$\frac{\partial S}{\partial c_i} = 2c_i - 2 \sum_{l=1}^{k} \mu_l d_{il}$$

$$= 0 \quad \text{for a minimum and therefore}$$

$$c_i = \sum_{l=1}^{k} \mu_l d_{il}. \tag{7}$$

To determine the μ_l we note that

$$\sum_i d_{i1} \sum_j \mu_j d_{ij} = 1,$$

$$\sum_i d_{il} \sum_j \mu_j d_{ij} = 0 \quad (l = 2, 3, \ldots, k),$$

that is

$$\sum_j \mu_j \sum_i d_{i1} d_{ij} = 1,$$

$$\sum_m \mu_m \sum_i d_{il} d_{im} = 0 \quad (l = 2, 3, \ldots, k). \tag{8}$$

This is a system of k equations in the unknowns $\mu_1, \mu_2, \ldots, \mu_k$. It is of interest to note that the system is most easily solved when the number of predictors omitted from the subset is low. This contrasts with the situation in which the subregression is computed directly from **R**. If this is done, then the amount of computation is low if the number of predictors omitted is high.

A special case of (6) arises when all predictors are to be included in the prediction. The solution of (8) is then:

$$c_i = \mu_1 d_{i1}$$

with $\mu_1 \Sigma d_{j1}^2 = 1$ and therefore

$$\mu_1 = 1/\Sigma d_{j1}^2, \quad c_i = d_{i1}/\Sigma d_{j1}^2$$

and the residual variance is

$$s^2 = \Sigma c_i^2 = 1/\Sigma d_{j1}^2. \tag{9}$$

We see that c_i is directly proportional to d_{i1}.

Thus the regression weights heavily the eigenvectors with large d_{i1}, and hence low s^2 values. As noted earlier, a low s^2 implies a low λ, and we may think of (9) as filling out and quantifying the discussion at the end of the preceding section. Equations (8) show how the **D** matrix may be used to find the conventional multiple regression, and also any subregression that is shown up by the analysis as promising.

4. Interpretation of the μ_l

Let us consider the following generalization of equations (8):

Minimize Σc_i^2,

subject to

$$\Sigma c_i d_{il} = r_l \quad (l = 1, 2, \ldots, k). \tag{10}$$

This system has the solution

$$c_i = \sum_{l=1}^{k} \mu_l d_{il},$$

with

$$\sum_{i=1}^{p} \sum_{l=1}^{k} \mu_l d_{il} d_{im} = r_m \quad (m = 1, 2, \ldots, k). \tag{11}$$

The residual variance is

$$\sum_{i=1}^{p} c_i^2 = \sum_{i=1}^{p} \left\{ \sum_{l=1}^{k} \mu_l d_{il} \right\}^2$$

$$= \sum_{i=1}^{p} \sum_{l=1}^{k} \mu_l \sum_{m=1}^{k} \mu_m d_{il} d_{im}$$

$$= \sum_{l=1}^{k} \mu_l r_l. \tag{12}$$

Now in our formulation $r_1 = 1$ and $r_l = 0$ $(l = 2, 3, \ldots, k)$. Thus $\Sigma c_i^2 = \mu_1$, and so μ_1 can be interpreted as the residual variance of the subregression.

Suppose we leave $r_1 r_2 \ldots r_{k-1}$ fixed at $1, 0, 0, \ldots, 0$ and vary r_k. Then equations (11) show that the μ_l will be linear functions of r_k.

Let

$$\mu_l(r_k) = \mu_l + a_l r_k \quad (l = 1, 2, \ldots, k),$$

where $\mu_l(0) = \mu_l$ is the μ value defined by the system (8). The residual variance (12) is then

$$S = \Sigma \mu_l(r_k) r_l = \mu_1(r_k) + \mu_k(r_k) r_k$$

$$S = \mu_1 + a_1 r_k + (\mu_k + a_k r_k) r_k$$

$$\partial S/\partial r_k = a_1 + \mu_k + 2a_k r_k$$

and S is minimized if

$$2a_k r_k = -(a_1+\mu_k). \tag{13}$$

Now if r_k assumes this value, then the constraint $\Sigma c_i d_{ik} = r_k$ does not constrain the optimum. Hence $\mu_k(r_k) = 0$, that is,

$$\mu_k - \tfrac{1}{2}(a_1+\mu_k) = 0 \tag{14}$$

or equivalently $a_1 = \mu_k$. Thus $S = \mu_1 + \mu_k r_k$.

Now if we allow the r_k to "float" so as to minimize S, then in fact we are including x_k in the regression, and the value assumed by r_k will be the coefficient of x_k in the subregression using $x_k, x_{k+1}, \ldots, x_p$. This implies that $\mu_k r_k$ is the reduction in residual variance that results if x_k is added to the subset of predictors.

Result 1

If row i of $\mathbf{D}$ is equated to zero and solved for x_1, the prediction so obtained has a residual variance of $1/d_{i1}^2$. This result is of some significance. If row i has non-zero loadings only on some subset of the predictors, then the prediction involves only this subset. Hence $1/d_{i1}^2$ provides an upper bound to the residual variance of the subregression using only these predictors. If d_{i1} is large, then this upper bound will be low.

Result 2 (*Substitution*)

Suppose rows i and j both have non-zero loadings on some predictor x_n. Then by solving row j for x_n and substituting in row i, we get a modified row i with loadings d_{im}^* defined by

$$d_{im}^* = d_{im} - r d_{jm},$$

where $r = d_{in}/d_{jn}$.

The variance of the modified row i is $(1+r^2)$; and, if solved for x_1, yields a residual variance of

$$(1+r^2)/(d_{i1}-rd_{j1})^2 = 1/d_{i1}^2\{(1+r^2)/(1-rd_{j1}/d_{i1})^2\}.$$

This provides an upper bound for the residual variance obtained by removing x_n from the subset. The bound is a minimum if $r = -d_{j1}/d_{i1}$, a fact that guides the choice of substitutions.

5. Rotation of D

The matrix $\mathbf{D}$ was derived from the eigenvectors of $\mathbf{R}$. The mathematics of equations (3) onwards, however, merely assumed that the y_i were uncorrelated and of unit variance. These properties are preserved if the rows of $\mathbf{D}$ are subjected to any orthogonal transformation. Let consider in what way we can improve the rows of $\mathbf{D}$.

(1) We have seen that if row i of $\mathbf{D}$ is solved for x_j, a residual variance of $1/d_{ij}^2$ results. This is low if d_{ij}^2 is large. Thus we would like to have as many large d_{ij}^2 as possible.

(2) Result 1 above gives useful information on subregressions if row i of $\mathbf{D}$ loads only on some subset of the prediction. Thus we would like the rotation to induce as many zeroes as possible into $\mathbf{D}$.

These criteria for rotation of $\mathbf{D}$ are very similar to the criteria used by factor analysts trying to produce simple structure in factor matrices. One approach leads to the use of the Varimax criterion for rotation. Varimax has the advantage over other possible criteria of being well known, and is included in many program packages. After rotation by Varimax, the rotated **D** matrix has many near zero elements and the remainder tend to be as large as possible. In what follows, we assume that **D** has been rotated to attain simple structure.

The residual variance is $1/\Sigma d_{j1}^2$. Now as a result of the rotation, most d_{i1} will be close to zero. Suppose only one d_{i1} is large and so dominates Σd_{j1}^2. Then row i is an essentially unique good predictor (cf. equation (9)). Any other good predictor will be obtained from this row by a substitution process. As a result of the rotation, it will generally not be possible by substitution to get a more parsimonious subregression than that implied by row i. Thus the major function of substitution is replacing predictors that are costly to measure with others.

If two of the d_{j1}—say d_{i1} and d_{n1}—are large, and the remainder near zero, then each implies a subregression using the predictors strongly identified with rows i and n. These subregressions have upper bounds $1/d_{i1}^2$ and $1/d_{n1}^2$ respectively. The rows may be combined to yield a subregression containing the union of the subsets involved in each row. The residual variance has upper bound $1/(d_{i1}^2+d_{n1}^2)$. Other good subsets may be implied by substitution. This discussion extends to the situation of several large d_{j1} in an obvious way.

6. Implementation

The use of **D** to suggest possible subregression is most valuable when any subregressions can be fitted immediately, and the residual variance and regression weights found. In practice, this is most easily achieved by means of an interactive remote access terminal into a computer. An interactive Fortran IV program has been written to accept the list of predictors to be omitted and compute the subregressions using (7) and (8). In this way, the user can identify possible subregressions from the rows of **D** and fit them immediately.

The use of **D** and the program is illustrated by means of the following examples.

Example 1

The data are from Stone (1945) and have been discussed by Kendall (1957). The variables are:

x_1 = annual consumption of beer in U.K.,

x_2 = real income,

x_3 = retail price of beer,

x_4 = cost of living index,

x_5 = strength of beer,

x_6 = time.

In Table 1 we give **R** and **D**. The value of Σd_{j1}^2 is 91·65 implying a residual variance of 0·0109. The fourth and sixth rows of **D** have large d_{i1}. Row 4 identifies with x_4, x_5 and x_6. $1/d_{41}^2 = 0{\cdot}047$, giving an approximate upper bound for the residual variance of this subregression. The actual residual variance is 0·032.

Row 6 identifies with x_3, x_4 and x_6 and $1/d_{61}^2 = 0{\cdot}015$ is an approximate upper bound for the residual variance of this subregression. The actual residual is 0·019.

Row 5 identifies a near multicollinearity involving x_2, x_4 and x_6. We may use it to substitute for x_6 in row 6 with little increase in residual variance. The actual residual on x_2, x_3 and x_4 is 0·0192.

TABLE 1

Data from Example 1

		Correlation matrix **R**			
1					
−·4539	1				
·0317	−·6091	1			
·8992	−·6564	·4477	1		
·6014	−·5093	−·2567	·3981	1	
−·7102	·9168	−·4628	−·8310	−·6496	1
		Rotated **D** *matrix*			
−·02	·00	·14	·96	·01	·02
1·43	−·03	−·47	−·98	−·05	−·05
·21	−·18	−·13	−1·60	·00	−1·63
−4·60	·14	·19	6·02	3·32	3·87
−·06	−3·81	−·94	1·66	·09	4·47
−8·27	−·50	−3·71	9·55	·72	1·25

In fitting, we find that x_2 has a low weight (as in fact some simple calculations would have suggested). This raises the possibility of a fit using only x_3 and x_4. The residual variance is 0·0194. An optimal combination of rows 4 and 6 yields an approximate upper bound of 0·011, with predictors x_3, x_4, x_5 and x_6. The actual residual is 0·011.

In Table 2, we give a summary of these results as provided by the interactive program. In each run, a number of predictors have been excluded. They are marked by having μ values, but no regression weights. The variables in the subregression have no μ values, but do have regression weights. The variable x_1 has both a weight (identically 1) and a μ value, the μ value being the residual variance of the subregression.

Incidentally, a scan of **R** suggests that x_3 may be a good candidate for elimination, having a correlation for only 0·03 with the dependent variable. The rows of **D**, however, do not suggest any way of eliminating x_3 except by the use of row 4 (bound = 0·047). That this conclusion is correct is confirmed by the interactive program. If x_3 is eliminated, the residual variance goes from 0·0109 to 0·0287, almost trebling. On the other hand, x_2 which is quite highly correlated with x_1 is seen from the rows of **D** to be readily eliminable. The residual variance is 0·011.

In this discussion, it is assumed that the user has no strong preference as to which predictor he would like to use. Suppose that for some reason he would prefer to avoid use of say x_4. A scan of the eigenvectors suggests no way in which x_4 may be eliminated cheaply. This impression is confirmed by the interactive program—the residual variance if x_4 is omitted is 0·241. On the other hand, x_6 can be substituted from row 5 and so eliminated at little cost.

Other possibilities may be suggested by the rows of **D**, and checked out using the interactive program.

TABLE 2

Runs from Example 1

	Subregressions					
Run	x_1	x_2	x_3	x_4	x_5	x_6
1						
μ_l	·0323	·0179	−·0698			
weights	1			−1·25	−·55	−·69
2						
μ_l	·0193	−·0042			·0180	
weights	1		·47	−1·09		·02
3						
μ_l	·0194	−·0098			·0417	−·0050
weights	1		·46	−1·11		
4						
μ_l	·0192				·0324	·0024
weights	1	·02	·47	−1·10		
5						
μ_l	·0287		−·0561			
weights	1	−·20		−1·16	−·50	−·39
6						
μ_l	·0110	−·0027				
weights	1		·31	−1·17	−·23	−·27
7						
μ_l	·0109					
weights	1	·04	·32	−1·18	−·23	−·31

Example 2

The second example is based on a mining problem. It is required to predict x_1, the yield of a washing plant from input (x_2 to x_6) and output (x_7 to x_{10}) characteristics of the feedstock. There are known near multicollinearities between x_4, x_5 and x_6, and between x_8, x_9 and x_{10}. There is also a known high correlation between x_3 and x_4, and between x_7 and x_8. The cost of measurement of the predictors is as follows:

very cheap: x_2, x_4 and x_8,

fairly cheap: x_5 and x_9

expensive: the remainder

and the predictor selected should, if possible, reflect these costs.

The correlation matrix **R** is given in Table 3, and the **D** matrix in Table 4. We notice that rows 7, 8, 9 and 10 highlight clearly the above-mentioned known interrelationships amongst the predictors. $\Sigma d_{j1}^2 = 6{\cdot}4$, while $d_{61}^2 = 6{\cdot}15$. Thus the residual variance is 0·156, and the multiple regression consists essentially of solving row 6 for x_1. This row identifies strongly with x_4, x_5, x_6 and x_8. Now row 10 may be used to eliminate the (costly) x_6. The residual on x_4, x_5 and x_8 is 0·187. In fitting the latter subregression, we find that the coefficient of x_5 is not large. This is to be expected from rows 6 and 10 of **D**. Thus we try eliminating x_5, and arrive at a subregression using only x_4 and x_8, and having residual variance 0·187.

Table 3

Data from Example 2

Correlation matrix **R**

1									
·162	1								
·881	·126	1							
−·883	−·123	−·998	1						
·213	·137	·234	−·233	1					
·751	·047	·854	−·856	−·304	1				
·413	·272	·610	−·609	·254	·460	1			
−·392	−·300	−·610	·612	−·176	−·504	·968			
·182	·121	·144	−·138	·546	−·157	·343	−·216		
·259	·206	·487	−·491	−·163	·568	·702	−·809	−·399	1

Table 4

Data from Example 2

Rotated **D** *matrix*

·00	·00	−·01	−·66	·03	·48	·00	·10	·00	−·09
·00	·00	·00	−·72	·72	−·74	·00	·14	−·16	·12
·00	·00	·00	·31	·00	−·29	·00	−·59	−·06	·69
−·35	1·10	−1·08	−3·21	−1·02	−1·67	·33	3·22	1·55	2·58
·00	·00	·00	−·39	·29	−·21	·00	·78	−·78	·51
2·48	−·16	·98	7·23	2·13	3·74	−·16	−1·48	−·69	−·67
−·09	·08	−·78	−1·37	−·41	−·63	4·81	5·81	·22	1·41
·11	−·06	22·37	18·41	−2·11	−4·16	−·16	−2·43	−1·59	−2·34
−·08	·10	−1·77	−4·00	−1·19	−2·18	·40	59·01	37·50	62·39
·32	−·06	1·31	77·80	41·30	77·66	−·10	−2·30	−1·18	−2·10

Table 5

Runs from Example 2

Subregressions

Run	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1										
μ_e	·156									
weights	1	−·13	·92	7·25	2·93	5·39	−·16	−1·73	−·91	−1·35
2										
μ_e	·181	·107	−·001				·014		·079	−·047
weights	1			6·75	3·09	5·84		−·246		
3										
μ_e	·187	·106	−·001			−·001	·014		·085	−·051
weights	1			1·02	−·02			−·235		
4										
μ_e	·187	·107	−·001		·016	−·009	·016		·093	−·056
weights	1			1·03				−·234		

7. Conclusion

These examples illustrate the use of the reduction to canonical form to investigate the interrelationships between x_1 and the predictors, and amongst the predictors. In both cases aspects of the data emerged that were not immediately apparent from **R**. The interrelationships found suggest possible subregressions which can then be fitted. This approach is believed to represent an improvement over existing standard techniques in that possible subregressions and substitutions can be identified easily without the heavy computation and post-computation analysis implied by the best subset method. In contrast with the stepwise and factor analysis methods, we find all good subregressions and alternatives implied by the interrelationships existing.

The technique requires program support for the computation of **R**, its reduction to canonical form and Varimax rotation of **D**. These are all standard operations easily handled by statistical library programs.

There is scope for research into the use of rotation criteria other than Varimax, and it is anticipated that a criterion reflecting structure requirements more accurately than Varimax will highlight alternative regressions better.

References

Beale, E. M. L., Kendall, M. G. and Mann, D. W. (1967). The discarding of variables in multivariate analysis. *Biometrika*, **54**, 357–366.

Browne, M. W. (1969). Factor analysis models and their application to prediction problems. Ph.D. Thesis, University of South Africa.

Daling, J. R. and Tamura, H. (1970). Use of orthogonal factors for selection of variables in a regression equation—an illustration. *Appl. Statist.*, **19**, 260–268.

Efroymson, M. A. (1965). Multiple regression analysis. In *Mathematical Methods for Digital Computers*, pp. 191–203. New York: Wiley.

Garside, M. J. (1971). Some computational procedures for the best subset problem. *Appl. Statist.*, **20**, 8–15.

Hotelling, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. *Brit. J. Statist. Psychol.*, **10**, 69–79.

Kendall, M. G. (1957). *A Course in Multivariate Analysis*. London: Griffin.

Massy, W. F. (1965). Principal component regression in exploratory statistical research. *J. Amer. Statist. Ass.*, **60**, 234–256.

Stone, R. (1945). The analysis of market demand. *J. R. Statist. Soc.* A, **108**, 286–382.