Paper 277

Stepwise Regressor Selection for Interdependence Analysis and Multivariate Multiple Regression

Lynette Duncan, University of Arkansas, Fayetteville, AR
James E. Dunn, University of Arkansas, Fayetteville, AR

ABSTRACT

The SWEEP operator in SAS/IML® software is used to solve two classes of multivariate problems, namely stepwise predictor selection in (1) *interdependence analysis* as defined by Beale, Kendall, and Mann (1967), and (2) *multivariate, multiple regression*. Though both problems have lain quiescent, they are resurrected here because of their relevance to analysis of massive data sets. In both cases, each forward selection step is followed by a "look back", where the latter involves switching 1-tuples, 2-tuples, or higher k-tuples of previously selected predictors. In this sense, both algorithms resemble SELECTION=MAXR in procedure REG, but with the capability of simultaneously switching more than single predictor variables in and out of the current model, and with the added important provision of a multivariate response. Examples of interdependence analysis are drawn from electrical engineering, yielding subsets of fundamental parameters which are used to characterize general purpose IC chips. Selection in multivariate, multiple regression is illustrated by two examples focusing on sources of variation for dioxin and furan yields in municipal waste incineration. The paper will be of interest to those who are proficient in regression-based model building techniques.

Key words: redundancy, Gummel-Poon, sweep, module, IML, MAXR

INTRODUCTION

In the following development, column vectors are indicated by bold, lower case letters, e.g., **x**, **y**, or Greek letters underscored, e.g., $\underline{\sigma}$; matrices are denoted by bold upper case letters, e.g., **A**, or capital Greek letters, e.g., Σ.

A consensus, if there was one, from a recent NRC workshop on analysis of massive data sets (1996), is that we will continue to use rather standard statistical techniques, but applied to samples drawn from these data sets rather than the whole. Implicit is the assumption that even our samples may be too large to store in memory as an $n \times v$ array. In some cases, this may be circumvented by use of recursive estimators, which require only storage of the previous estimator and the current observation. In particular, regression analysis, as well as many familiar multivariate techniques, such as principal components analysis, requires only the estimated $v \times v$ SS & CP matrix, perhaps rescaled as a correlation matrix in order to standardize the variables, or simply to provide a better conditioned matrix. We recall the form of this recursion, for those not familiar with recursive techniques: If

$$\mathbf{C}_n = \sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}}_n)(\mathbf{y}_i - \bar{\mathbf{y}}_n)'$$

represents the SS & CP matrix based on the first n observations, where

$$\bar{\mathbf{y}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_i$$

is their centroid, then with an additional observation, $\mathbf{y}_{n+1}$

$$\mathbf{C}_{n+1} = \mathbf{C}_n + \frac{n}{n+1}(\mathbf{y}_{n+1} - \bar{\mathbf{y}}_n)(\mathbf{y}_{n+1} - \bar{\mathbf{y}}_n)',$$

and $$\bar{\mathbf{y}}_{n+1} = \frac{n}{n+1}\bar{\mathbf{y}}_n + \frac{1}{n+1}\bar{\mathbf{y}}_{n+1}.$$

Model building, based on sequential predictor selection in a regression model, is a major tool familiar to all applied statisticians, e.g., the SELECTION= options in PROC REG. Our extension here is to that of sequential predictor selection for a multivariate response. Univariate t-tests for the partial effects of predictors are replaced by F-tests based on Hotelling's T^2. Our algorithm most closely resembles that of SELECTION=MAXR, alternating a forward selection step followed by a "look back" based on predictor switching. However, in our case, we have not limited ourselves to examining the effects of all possible 1-tuple switches between predictors in and out of the model, but rather have extended our capability to include all possible 2-tuple, 3-tuple, … switches.

Interdependence analysis, a term coined by Beale, *et al.*, in 1967 reappeared once in the title of a monograph by Boyce, *et al.*, in 1974, and then seems to have disappeared. The idea is simple and direct: Given a multivariate set of responses (and no predictors), find a subset which will predict the others while satisfying some optimality criterion. Like principal components and factor analysis, isolation of fundamental variables is the goal, but in this case, no new, additional latent variables are created. An application which immediately comes to mind is the continual search for surrogate compounds to predict EPA's extensive list of hazardous organics in the environment. An alternative application related to properties of multi-transistor chips was called to our attention by a graduate student in electrical engineering (Garimella, 1997).

Beale, *op.cit.* focused on the criterion of maximizing the minimum squared multiple correlation (R^2). However, like procedures TRANSREG and CANCORR, we have focused alternatively on maximizing average squared multiple correlation, i.e., the *redundancy*, while continuing to monitor minimum R^2 at each step.

Even though CEIR, Ltd. software apparently existed at one time to implement interdependence analysis (Beale, *op.cit.*), we have not determined its fate. A major part of our interest was to exercise the capability of the PROC IML SWEEP operator, as well as the general indexing capabilities of PROC IML, in order to produce answers to questions posed in terms of actual data sets. In that we anticipate many problems for which v>>n, we have not implemented a branch-and-bound step, so popular in older sequential selection algorithms. To see a citation of 7 minutes for analysis of 16 variables (Beale, *op.cit.*) now seems incomprehensible. By comparison, we find our execution of macros and modules, built around SWEEP, to be very quick, in spite of the fact that we fit some possibly redundant models.

The following definitions are general where it is presumed that $[\mathbf{x}' : \mathbf{y}']$ represents a p + q dimensional vector of random variables whose values are represented by the data.

Defn. **P** (read "capital rho") is a *correlation matrix.* $\{\mathbf{P}\}_{ij} = \rho_{ij}$ is the simple correlation between the i^{th} and the j^{th} variables, and **P** is symmetric.

Defn. $\rho^2_{k \bullet \mathbf{x}}$ is the *squared multiple correlation* between y_k and a predictor set $\mathbf{x}' = \{x_1, \ldots, x_p\}$, k = 1,…,q and reflects the fraction of the variation in y_k which may be attributed to variation in **x**.

Defn. $\rho_{jk \bullet \mathbf{x}}$ is the *partial correlation* between y_j and y_k, conditional on holding **x** constant. We infer that the effect of **x** has been "partialled" out of the association between y_j and y_k.

Defn. Given $p \times 1$ **x** and $q \times 1$ **y**, *redundancy* is defined as

$$\Theta(\mathbf{x}) = \sum_{j=1}^{q} \rho^2_{j \bullet \mathbf{x}} / q,$$

the average squared multiple correlation which results when each element of **y** is predicted by the best linear function of elements of **x**.

Defn. Given only p and q, we seek optimal **x** corresponding

to $\Theta_{max} = \max_{\mathbf{x}} \Theta(\mathbf{x})$, that set of predictors yielding maximum redundancy for prediction of all elements of **y**.

INTERDEPENDENCE ANALYSIS

COMPUTATIONAL ALGORITHM FOR FORWARD SELECTION

At each iterative step, let s = number of variables in **x** and t = number of variables in **y**. Initially, s = 0 and t = p + q, so that $\mathbf{P} = \mathbf{P}_{yy}$.

1. **If s = 0:**
   a. Compute $v_k = \sum_{j\neq k}^{t} \rho_{kj}^2 = \sum_{j=1}^{t} \rho_{kj}^2 - 1$ for each row of $\mathbf{P}_{yy}$.
   b. Transfer y_i to **x** if $v_i = \max\{v_1,\ldots,v_t\}$, and set s = 1 and t = t - 1.
   c. Compute the *redundancy,* $\Theta(\mathbf{x}) = v_i/t$.
   d. Compute $\Psi_{y(x)} = \mathbf{P}_{xy}'\mathbf{P}_{xx}^{-1}\mathbf{P}_{xy}$, using PROC IML SWEEP.

   If s > 0:
   a. Partition the *correlation matrix*, $\mathbf{P} = \begin{bmatrix} \mathbf{P}_{xx} & \mathbf{P}_{xy} \\ \mathbf{P}_{xy}' & \mathbf{P}_{yy} \end{bmatrix} \begin{matrix} (s) \\ (t) \end{matrix}$ (s) (t).
   b. Compute the *conditional covariance matrix of standardized variables*, $\Sigma_{\mathbf{y}_S|\mathbf{x}_S} = \mathbf{P}_{yy} - \mathbf{P}_{xy}'\mathbf{P}_{xx}^{-1}\mathbf{P}_{xy} = \mathbf{A}$.
   c. Define $\mathbf{d} = (a_{11},\ldots,a_{tt})'$, written as a column vector, and $\mathbf{D} = \text{diag}(\sqrt{a_{11}},\ldots,\sqrt{a_{tt}})$, where $a_{kk} = 1 - \rho_{k\bullet\mathbf{x}}^2$.
   d. Compute the *partial correlation matrix*, $\mathbf{P}_{\mathbf{y}|\mathbf{x}} = \mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}$, with (j,k)th element $\rho_{jk\bullet\mathbf{x}}$, the *conditional correlation* between y_j and y_k given **x**.
   e. Compute $\mathbf{B} = \mathbf{J} - \mathbf{P}_{\mathbf{y}|\mathbf{x}}^{\#\#2}$, where **J** is a t × t matrix of 1's, and ##2 indicates that the elements of $\mathbf{P}_{\mathbf{y}|\mathbf{x}}$ are squared.
   f. Compute $\mathbf{f} = t^{-1}\mathbf{B}\mathbf{d}$. Note that $f_k = \sum_{j\neq k}^{t}(1-\rho_{j\bullet\mathbf{x}}^2)(1-\rho_{jk\bullet\mathbf{x}}^2)/t$.
   g. Transfer y_i to **x** if $f_i = \min\{f_1,\ldots,f_t\}$.
   h. Set s = s + 1 and t = t - 1, compute $\Psi_{y(x)}$ based on a newly defined partition of **P**, and the *redundancy*, $\Theta(\mathbf{x}) = \text{tr}(\Psi_{y(x)})/t$.

2. Based on the new partition of **P**, compute the conditional covariance matrix $\Sigma_{\mathbf{y}_S|\mathbf{x}_S}$, *generalized variances*, $\Delta = |\Sigma_{\mathbf{y}_S|\mathbf{x}_S}|$ and $\delta = |\mathbf{P}_{xx}|$, and $\rho^2(\text{min}) = \min \text{diag}(\Psi_{y(x)}) = \min\{\rho_{1\bullet\mathbf{x}}^2,\ldots,\rho_{t\bullet\mathbf{x}}^2\}$.
3. Compute the p-value associated with a multivariate test of significance of the last predictor added to **x**.
4. Repeat steps 1 to 3 until a satisfactory solution is attained, as reflected by the sequences $\{\Theta_1,\Theta_2,\ldots\}$, $\{\Delta_1,\Delta_2,\ldots\}$, $\{\delta_1,\delta_2,\ldots\}$, $\{\rho_1^2(\text{min}),\rho_2^2(\text{min}),\ldots\}$, and $\{p_1,p_2,\ldots\}$, where the subscripts index the iterative steps. Optimally $\rho^2(\text{min}) \to 1$, $\Theta \to 1$, $\Delta \to 0$, $\delta \equiv 1$, and p should be compared to a typical, user-supplied significance level, α.

OPTIMAL PROPERTIES OF THE ALGORITHM

Result 1. For s = 1, the algorithm is optimal.

Pf. $\mathbf{P}_{xy} = [\rho_{x1},\ldots,\rho_{xq}]$, where $\rho_{xj} = \text{corr}[x,y_j]$, and $\mathbf{P}_{xx}=1$

$\Rightarrow$ $tr(\mathbf{P}_{xy}'\mathbf{P}_{xx}^{-1}\mathbf{P}_{xy}) = \sum_{j=1}^{t-1}\rho_{xj}^2$. Thus, $\Theta_{max} = \max_x\{\sum_{j=1}^{t-1}\rho_{xj}^2\}/(t-1)$.

Result 2. Conditional on s-1 dimensional x, Θ_s is optimal for s=2,3,….

Pf. For any p random variables, Kendall and Stuart (1979) derive a result relating a squared multiple correlation to a sequence of partial correlations involving those same variables, namely

$1-\rho_{1\bullet 2\ldots p}^2 = (1-\rho_{12}^2)(1-\rho_{13\bullet 2}^2)(1-\rho_{14\bullet 23}^2)\ldots(1-\rho_{1p\bullet 23\ldots(p-1)}^2)$,

from which we infer the recursion

$1-\rho_{1\bullet 2\ldots p}^2 = (1-\rho_{1\bullet 2\ldots(p-1)}^2)(1-\rho_{1p\bullet 23\ldots(p-1)}^2)$.

In the context of our problem, for fixed x and indices j and k referring to elements y_j and y_k of **y**, this implies

$1-\rho_{j\bullet\mathbf{x}k}^2 = (1-\rho_{j\bullet\mathbf{x}}^2)(1-\rho_{jk\bullet\mathbf{x}}^2)$.

Choice of y_i to maximize the *redundancy* $\sum_{j\neq k}^{t}\rho_{j\bullet\mathbf{x}k}^2/(t-1)$, over the integer set $k \in \{1,\ldots,t\}$, is equivalent to determining y_i such that $\sum_{j\neq i}^{t}(1-\rho_{j\bullet\mathbf{x}i}^2) < \sum_{j\neq k}^{t}(1-\rho_{j\bullet\mathbf{x}k}^2)$ for every $k \neq i$ which, from the recursion, is equivalent to $\sum_{j\neq i}^{t}(1-\rho_{j\bullet\mathbf{x}}^2)(1-\rho_{ji\bullet\mathbf{x}}^2) < \sum_{j\neq k}^{t}(1-\rho_{j\bullet\mathbf{x}}^2)(1-\rho_{jk\bullet\mathbf{x}}^2)$ for every k≠i, the latter criterion being specified by steps 1f and 1g of the algorithm.

Remark: Even though this sequential solution may not be the global solution at any particular step, there can exist no better solution without replacing x-variables which were selected in previous steps. The computational algorithm is extremely efficient in that $\rho_{j\bullet\mathbf{x}}^2$ results as an intermediate step in obtaining $\rho_{jk\bullet\mathbf{x}}^2$.

BACK-SWITCHING

Given p-dimensional **x** and q-dimensional **y** and their associated redundancy $\Theta(\mathbf{x})$, consider 1-1 exchanges of 1-tuples, 2-tuples, 3-tuples, etc. of the elements of **x** and **y** until $\max \Theta(\mathbf{x}|s)$ is attained. Most forward selection steps result in $\Theta(\mathbf{x}) = \max \Theta(\mathbf{x}|s)$, but exceptions occur with sufficient frequency as to require further checking. Such an algorithm is the basis of SELECTION=MAXR, though only 1-1 1-tuple switching of predictors of a univariate response is implemented there.

An efficient means of evaluating the redundancy of a multivariate response, **y**, given the predictor set, **x**, is by means of the PROC IML SWEEP operator. In general, if $\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{bmatrix}$ is swept with respect to **x**, then $\begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-1} & (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ -\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \end{bmatrix}$ results. If we start with partitions of a correlation matrix, $\begin{bmatrix} \mathbf{P}_{xx} & \mathbf{P}_{xy} \\ \mathbf{P}_{xy}' & \mathbf{P}_{yy} \end{bmatrix}$ and this is swept with respect to **x**, then

$$S = \begin{bmatrix} \mathbf{P}_{xx}^{-1} & B = \mathbf{P}_{xx}^{-1}\mathbf{P}_{xy} \\ -B' & \mathbf{P}_{yy} - \mathbf{P}_{xy}'\mathbf{P}_{xx}^{-1}\mathbf{P}_{xy} \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ -S_{12}' & S_{22} \end{bmatrix} \qquad (1)$$

results, where $B = \mathbf{P}_{xx}^{-1}\mathbf{P}_{xy}$ is a matrix of *standard regression coefficients*. In particular, the (2,2) partition of the result is $S_{22} = \mathbf{P}_{yy} - \mathbf{P}_{xy}'\mathbf{P}_{xx}^{-1}\mathbf{P}_{xy}$, with the set $\{1-\rho_{j\bullet\mathbf{x}}^2\}$ appearing as its principal diagonal, so that $\Theta(\mathbf{x}) = 1 - \text{tr}(S_{22})/t = \sum_{j=1}^{q}\rho_{j\bullet\mathbf{x}}^2/t$.

Rather than re-shaping the correlation matrices so that $\mathbf{P}_{xx}$ always appears in the upper, left hand corner in preparation for SWEEP, it obviously is faster simply to SWEEP the original correlation matrix using the current contents of IX as the index set of **x**, based on use of the LOC command. If S is the result of the SWEEP with respect to IX, and IY is the index set of the remaining y's, then $\Theta(\mathbf{x}) = 1 - \text{TRACE}(S[\text{IY},\text{IY}])/t$ completes the redundancy evaluation.

As a closing remark, ideally the maximum order of back-switching should be an option available to you. Currently 1-tuple switching requires two DO loops; 2-tuple switching

requires four DO loops; etc. We are still working to develop a more efficient indexing method which avoids compounding the number of DO loops.

F-TEST FOR LAST ADDED PREDICTOR

Even though it is possible to invoke procedure GLM in MANOVA mode in order to test the significance of all regression coefficients associated with the last added predictor (or any other predictor, for that matter), the necessity of alternating between calls to PROCs IML and GLM is disruptive to the flow of successive model-building steps. The following development shows that PROC IML SWEEP can be used to obtain a test statistic in the form of Hotelling's T^2. To the extent that multi-normality holds, an exact F-test results.

Consider a test of the general *multivariate* linear hypothesis,

H_0: $\mathbf{GB} = \varnothing$,

where **B** is p × q, **G** is m × p, and rank(**G**) = m. **H** and **E** matrices, in terms of the estimator $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, are respectively

$$\mathbf{H} = \hat{\mathbf{B}}'\mathbf{G}'[\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}']^{-1}\mathbf{G}\hat{\mathbf{B}} = \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}'[\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}']^{-1}\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

$$\mathbf{E} = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Assuming that both **X** and **Y** have their respective centroids adjusted out, and defining $\mathbf{D}_x$ and $\mathbf{D}_y$ to be diagonal matrices whose diagonal elements are square roots of principal diagonal elements of **X'X** and **Y'Y** respectively, then

$$\mathbf{D}_y^{-1}\mathbf{H}\mathbf{D}_y^{-1} = \mathbf{D}_y^{-1}\mathbf{Y}'\mathbf{X}\mathbf{D}_x^{-1}(\mathbf{D}_x^{-1}\mathbf{X}'\mathbf{X}\mathbf{D}_x^{-1})^{-1}\mathbf{D}_x^{-1}\mathbf{G}'$$
$$\bullet[\mathbf{G}\mathbf{D}_x^{-1}(\mathbf{D}_x^{-1}\mathbf{X}'\mathbf{X}\mathbf{D}_x^{-1})^{-1}\mathbf{D}_x^{-1}\mathbf{G}']^{-1}\mathbf{G}\mathbf{D}_x^{-1}(\mathbf{D}_x^{-1}\mathbf{X}'\mathbf{X}\mathbf{D}_x^{-1})^{-1}\mathbf{D}_x^{-1}\mathbf{X}'\mathbf{Y}\mathbf{D}_y^{-1}$$
$$= \hat{\mathbf{P}}_{xy}'\hat{\mathbf{P}}_{xx}^{-1}\mathbf{D}_x^{-1}\mathbf{G}'[\mathbf{G}\mathbf{D}_x^{-1}\hat{\mathbf{P}}_{xx}^{-1}\mathbf{D}_x^{-1}\mathbf{G}']^{-1}\mathbf{G}\mathbf{D}_x^{-1}\hat{\mathbf{P}}_{xx}^{-1}\hat{\mathbf{P}}_{xy}, \qquad (2)$$

$$\mathbf{D}_y^{-1}\mathbf{E}\mathbf{D}_y^{-1} = \mathbf{D}_y^{-1}\mathbf{Y}'\mathbf{Y}\mathbf{D}_y^{-1} - \mathbf{D}_y^{-1}\mathbf{Y}'\mathbf{X}\mathbf{D}_x^{-1}(\mathbf{D}_x^{-1}\mathbf{X}'\mathbf{X}\mathbf{D}_x^{-1})^{-1}\mathbf{D}_x^{-1}\mathbf{X}'\mathbf{Y}\mathbf{D}_y^{-1}$$
$$= \hat{\mathbf{P}}_{yy} - \hat{\mathbf{P}}_{xy}'\hat{\mathbf{P}}_{xx}^{-1}\hat{\mathbf{P}}_{xy}. \qquad (3)$$

Suppose that the effect of the last-added x, say x_p, is to be tested (or that of any other single predictor), so that 1 × p **G** takes the form

$$\mathbf{G} = [0 \ldots 0\ 1],$$

$$\mathbf{G}\mathbf{D}_x^{-1} = [0 \ldots 0\ \{\mathbf{D}_x\}_{pp}^{-1}],$$

$$\mathbf{G}\mathbf{D}_x^{-1}\hat{\mathbf{P}}_{xx}^{-1}\mathbf{D}_x^{-1}\mathbf{G}' = \{\hat{\mathbf{P}}_{xx}^{-1}\}_{pp}\Big/\{\mathbf{D}_x\}_{pp}^2$$

$$\Rightarrow [\mathbf{G}\mathbf{D}_x^{-1}\hat{\mathbf{P}}_{xx}^{-1}\mathbf{D}_x^{-1}\mathbf{G}']^{-1} = \{\mathbf{D}_x\}_{pp}^2\Big/\{\hat{\mathbf{P}}_{xx}^{-1}\}_{pp} = \{\mathbf{D}_x\}_{pp}^2(1-\hat{\rho}^2_{p\bullet 1,\ldots,p-1}), \qquad (4)$$

using the general property that the reciprocal of the k^{th} principal diagonal element of $\hat{\mathbf{P}}_{xx}^{-1}$ defines 1 minus the squared multiple correlation between the k^{th} x and the remaining x-variables in the set. Substituting equation (4) into equation (2) yields

$$\mathbf{D}_y^{-1}\mathbf{H}\mathbf{D}_y^{-1} = (1-\hat{\rho}^2_{p\bullet 1,\ldots,p-1})\hat{\mathbf{P}}_{xy}'\hat{\mathbf{P}}_{xx}^{-1}\begin{bmatrix}0\\ \vdots\\ 0\\ 1\end{bmatrix}\begin{bmatrix}0 & \cdots & 0 & 1\end{bmatrix}\hat{\mathbf{P}}_{xx}^{-1}\hat{\mathbf{P}}_{xy}$$

$$= (1-\hat{\rho}^2_{p\bullet 1,\ldots,p-1})\underline{\upsilon}_p\underline{\upsilon}_p',$$

where $\underline{\upsilon}_p = \hat{\mathbf{P}}_{xy}'\hat{\mathbf{P}}_{xx}^{-1}\begin{bmatrix}0\\ \vdots\\ 0\\ 1\end{bmatrix} = B'\begin{bmatrix}0\\ \vdots\\ 0\\ 1\end{bmatrix}$ contains the p^{th} (or any other) set of standard regression coefficients defined in expression (1).

Applying Roy's Union-Intersection principle,

$$\max \text{ch}[\mathbf{E}^{-1}\mathbf{H}] = \max \text{ch}[\mathbf{D}_y\mathbf{E}^{-1}\mathbf{D}_y\mathbf{D}_y^{-1}\mathbf{H}\mathbf{D}_y^{-1}]$$
$$= \max \text{ch}[(\mathbf{D}_y^{-1}\mathbf{E}\mathbf{D}_y^{-1})^{-1}\mathbf{D}_y^{-1}\mathbf{H}\mathbf{D}_y^{-1}]$$
$$= \max \text{ch}[(\hat{\mathbf{P}}_{yy} - \hat{\mathbf{P}}_{xy}'\hat{\mathbf{P}}_{xx}^{-1}\hat{\mathbf{P}}_{xy})^{-1}(1-\hat{\rho}^2_{p\bullet 1,\ldots,(p-1)})\underline{\upsilon}_p\underline{\upsilon}_p']$$
$$= (1-\hat{\rho}^2_{p\bullet 1,\ldots,(p-1)})\underline{\upsilon}_p'(\hat{\mathbf{P}}_{yy} - \hat{\mathbf{P}}_{xy}'\hat{\mathbf{P}}_{xx}^{-1}\hat{\mathbf{P}}_{xy})^{-1}\underline{\upsilon}_p$$
$$= \mathbf{T}^2/(n-p-1).$$

Alternatively, it is convenient to SWEEP

$$\begin{bmatrix}\hat{\mathbf{P}}_{yy} - \hat{\mathbf{P}}_{xy}'\hat{\mathbf{P}}_{xx}^{-1}\hat{\mathbf{P}}_{xy} & \underline{\upsilon}_p\\ \underline{\upsilon}_p' & 0\end{bmatrix}, \text{ yielding}$$

$$F = \frac{(n-p-q)}{q}\underline{\upsilon}_p'(\hat{\mathbf{P}}_{yy} - \hat{\mathbf{P}}_{xy}'\hat{\mathbf{P}}_{xx}^{-1}\hat{\mathbf{P}}_{xy})^{-1}\underline{\upsilon}_p\Big/\{\hat{\mathbf{P}}_{xx}^{-1}\}_{pp},$$

which is distributed as $F_{(q,n-p-q)}$ when H_0 is true.

Remark: Once the SWEEP with respect to **x** has been performed, the partial effect of *any* predictor in the x-set could be tested similarly by picking the appropriate row of *B* (to redefine $\underline{\upsilon}$) and the appropriate diagonal element of $\hat{\mathbf{P}}_{xx}^{-1}$.

Example 1. Divekar, *et al.* (1977) reported on the analysis of Gummel-Poon parameters, used to characterize a sample of n=35 general purpose IC chips comprised of various configurations of bipolar transistors. The 18 parameters were divided into mutually exclusive subsets of size 5 which depend on collector doping, and 13 which depend on emitter and base dopings. Only the estimated correlation matrices appeared in the article, and these are reproduced below as $\hat{\mathbf{P}}_1$ and $\hat{\mathbf{P}}_2$, respectively:

$$\hat{\mathbf{P}}_1 = \begin{array}{c} \begin{matrix} R_C & I_k & C_{CCB} & m_{CB} & \phi_{CB} \end{matrix} \\ \begin{bmatrix} 1 & -.4 & -.7 & .45 & .75 \\ & 1 & .47 & -.21 & -.6 \\ & & 1 & -.67 & -.56 \\ & & & 1 & .59 \\ & & & & 1 \end{bmatrix} \end{array}$$

$$\hat{\mathbf{P}}_2 = \begin{array}{c} \begin{matrix} \beta_F & \beta_R & R_{PB} & f_T & I_S & R_E & n_e & R_B & V_A & C_2 & C_{OEB} & \phi_{EB} & m_{EB} \end{matrix} \\ \begin{bmatrix} 1 & .73 & .26 & .34 & .55 & -.15 & -.06 & .81 & -.6 & .01 & .22 & -.13 & .6 \\ & 1 & .76 & .57 & .9 & .38 & .32 & .57 & -.88 & .51 & -.24 & .25 & .66 \\ & & 1 & .56 & .83 & .8 & .67 & .2 & -.7 & .71 & -.7 & .62 & .31 \\ & & & 1 & .6 & .4 & .44 & .28 & -.54 & .43 & -.55 & .41 & .28 \\ & & & & 1 & .54 & .58 & .56 & -.81 & .77 & -.44 & .35 & .59 \\ & & & & & 1 & .66 & -.2 & -.38 & .64 & -.78 & .68 & .04 \\ & & & & & & 1 & .11 & -.25 & .82 & -.72 & .66 & .04 \\ & & & & & & & 1 & -.42 & .23 & .15 & -.22 & .55 \\ & & & & & & & & 1 & -.46 & .23 & -.2 & -.56 \\ & & & & & & & & & 1 & -.6 & .45 & .26 \\ & & & & & & & & & & 1 & -.7 & -.07 \\ & & & & & & & & & & & 1 & -.21 \\ & & & & & & & & & & & & 1 \end{bmatrix} \end{array}$$

The question posed in each case is that of searching for the best subset of variables to be used as predictors for the others. Factor analysis was proposed as the appropriate methodology, and continues to be advocated in this context. R_C was chosen in the first case, while a 2-factor solution in the second case identified I_S and β_F, and possibly R_{PB}, as the controlling parameters. In both cases, interdependence analysis leads to different predictor subsets than those proposed by the original authors, as the following computational sequences illustrate. We have no way of knowing if the elements of these matrices were rounded for publication purposes, but if so, this might explain the discrepancies in our results. Maximum redundancy was used as the selection criterion for both subsets, allowing 1-tuple switching.

Subset-1

Selected Predictors	Redundancy	Min R^2	p-value	Comments
R_C	0.354	0.160	5.47E-9	Divekar, et al.

ϕ_{CB}	0.396	0.314	3.15E-10	
ϕ_{CB}, C_{CCB}	0.526	0.386	5.26E-7	
ϕ_{CB},C_{CCB}, I_k	0.641	0.588	5.09E-3	
(ϕ_{CB},C_{CCB}, I_k, m_{CB})	0.762	0.762	6.57E-3	Replacement offered
I_k,C_{CCB}, m_{CB}, R_C,	0.804	0.804	2.53E-6	1-tuple switch

ϕ_{CB} is a better surrogate than R_C for four others by either the maximum redundancy or minimum R^2 criterion. The final switch in response-predictor roles of ϕ_{CB} and R_C is typical of interdependence analysis, we believe, when it is allowed to run to completion. ϕ_{CB} was selected as the first predictor because it was most strongly related to the other four responses. By the same argument, it was selected as the response to be predicted in the last step because it was strongly related to all four predictors.

Subset-2

Selected Predictors	Redundancy	Min R^2	p-value	Comments
I_S	0.418	0.122	1.13E-12	
(I_S, C_{OEB})	0.609	0.393	3.50E-7	Replacement offered
C_{OEB}, β_R	0.614	0.412	1.78E-13	1-tuple switch
β_F,I_S	0.595	0.272	3.20E-11	Divekar, et al.
(C_{OEB},β_R, R_B)	0.671	0.481	6.26E-5	Replacement offered
β_R,R_B, n_e	0.688	0.400	1.27E-6	1-tuple switch
β_F,I_S,R_{PB}	0.658	0.376	8.78E-6	Divekar, et al.
(β_R,R_B,n_e, C_{OEB})	0.732	0.521	1.55E-4	Replacement offered
β_R,R_B,C_{OEB} C_2	0.735	0.486	5.59E-7	1-tuple switch
β_R,R_B, C_{OEB},C_2, m_{EB}	0.794	0.546	0.0116	
β_R,R_B, C_2, C_{OEB}, m_{EB}, f_T	0.835	0.691	0.309	

The final model selected contained 5 surrogate predictors, β_R, R_B, C_{OEB}, C_2, m_{EB}, for predicting 8 remaining responses, β_F, R_{PB}, f_T, I_S, R_E, n_e, V_A, and ϕ_{EB}. Optimal 2-tuple and 3-tuple subsets shared no predictors in common with the subsets selected by Divekar, *et al.* In particular, their solutions were characterized by depressed values for minimum R^2. The selection process was terminated with the 5-predictor model because addition of the best sixth predictor resulted in a non-significant p-value. P-values here and in the other examples refer to the last predictor or predictors in each list. However, the best 6-predictor model may have merit because of its marked improvement in minimum R^2.

This example is important in two respects: (a) It motivated re-examination of alternatives to factor analysis in the context of the problem, most recently by Garimella (1997), a recent EE student at the University of Arkansas, and (b) It re-emphasizes the fact that the correlation matrix alone, in absence of the original data, is sufficient to support useful multivariate regression algorithms. This should have important overtones in the context of analysis of *massive data sets*, where it is a given that it is impossible to store the entire data set in memory for statistical processing.

PREDICTOR SELECTION IN MULTIVARIATE, MULTIPLE REGRESSION

Suppose that $\mathbf{y}' = [y_1,\ldots,y_q]$ represent q, *a priori* specified response variables, to be predicted by a subset of *a priori* specified available predictors. At any particular forward selection step, let p-dimensional $\mathbf{x}$ represent predictors already selected in previous steps, and t-dimensional $\mathbf{z}$ represent potential predictors yet remaining.

FORWARD SELECTION

Using redundancy as the measure, choose z_i over z_j as a predictor, in addition to previously selected $\mathbf{x}$, if

$$\sum_{k=1}^{q}\rho^2_{k\bullet i\mathbf{x}} > \sum_{k=1}^{q}\rho^2_{k\bullet j\mathbf{x}},$$

or if $$\sum_{k=1}^{q}(1-\rho^2_{k\bullet i\mathbf{x}}) < \sum_{k=1}^{q}(1-\rho^2_{k\bullet j\mathbf{x}}),$$

or if $$\sum_{k=1}^{q}(1-\rho^2_{k\bullet\mathbf{x}})(1-\rho^2_{ki\bullet\mathbf{x}}) < \sum_{k=1}^{q}(1-\rho^2_{k\bullet\mathbf{x}})(1-\rho^2_{kj\bullet\mathbf{x}}),$$

based on the fundamental identity found in Kendall and Stuart (1979).

In order to implement these comparisons, suppose that we SWEEP

$$\left[\begin{array}{c|cc} \mathbf{P}_{xx} & \mathbf{P}_{xz} & \mathbf{P}_{xy} \\ \hline & \mathbf{P}_{zz} & \mathbf{P}_{zy} \\ (\text{sym.}) & & \mathbf{P}_{yy} \end{array}\right]$$

with respect to $\mathbf{x}$, resulting in

$$\mathbf{S} = \left[\begin{array}{c|c} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \hline \mathbf{S}_{12}' & \mathbf{S}_{22} \end{array}\right],$$

where

$$\mathbf{S}_{22} = \begin{bmatrix} \mathbf{P}_{zz} & \mathbf{P}_{zy} \\ \mathbf{P}_{zy}' & \mathbf{P}_{yy} \end{bmatrix} - \begin{bmatrix} \mathbf{P}_{xz}' \\ \mathbf{P}_{xy}' \end{bmatrix}\mathbf{P}_{xx}^{-1}\begin{bmatrix}\mathbf{P}_{xz} & \mathbf{P}_{xy}\end{bmatrix} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{12}' & \mathbf{M}_{22} \end{bmatrix},$$

whose principal diagonal elements may be represented as

$$\mathbf{d} = \text{vecdiag}(\mathbf{S}_{22}) = \begin{bmatrix} 1-\rho^2_{z_1\bullet\mathbf{x}} \\ \vdots \\ 1-\rho^2_{z_t\bullet\mathbf{x}} \\ \hline 1-\rho^2_{y_1\bullet\mathbf{x}} \\ \vdots \\ 1-\rho^2_{y_q\bullet\mathbf{x}} \end{bmatrix} = \begin{bmatrix} \mathbf{d}_1 \\ \hline \mathbf{d}_2 \end{bmatrix}.$$

The partial correlation matrix for $\mathbf{z}$ and $\mathbf{y}$, given $\mathbf{x}$, is $\mathbf{D}_{\sqrt{\mathbf{d}}}^{-1}\mathbf{S}_{22}\mathbf{D}_{\sqrt{\mathbf{d}}}^{-1}$, in an obvious notation. In particular, partial correlations between $\mathbf{z}$ and $\mathbf{y}$ appear in the upper, right-hand partition of this matrix in the form

$$\text{corr}(\mathbf{z},\mathbf{y}'|\mathbf{x}) = \mathbf{P}_{\mathbf{z},\mathbf{y}|\mathbf{x}} = \mathbf{D}_{\sqrt{\mathbf{d}_1}}^{-1}\mathbf{M}_{12}\mathbf{D}_{\sqrt{\mathbf{d}_2}}^{-1}.$$

Complementing squares of the elements of this matrix to 1 yields

$$\mathbf{H} = \mathbf{j}_{(t)}\mathbf{j}_{(q)}' - \mathbf{P}_{\mathbf{z},\mathbf{y}|\mathbf{x}}^{\#\#2} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{12}' & \mathbf{H}_{22} \end{bmatrix},$$

where

$$\mathbf{H}_{12} = \begin{bmatrix} 1-\rho^2_{z_1y_1\bullet\mathbf{x}} & \cdots & 1-\rho^2_{z_1y_q\bullet\mathbf{x}} \\ \vdots & & \vdots \\ 1-\rho^2_{z_ty_1\bullet\mathbf{x}} & \cdots & 1-\rho^2_{z_ty_q\bullet\mathbf{x}} \end{bmatrix}.$$

As a final step, if we compute $\mathbf{H}_{12}\mathbf{d}_2$, then

$$\{\mathbf{H}_{12}\mathbf{d}_2\}_j = \sum_{k=1}^{q}(1-\rho^2_{y_k\bullet\mathbf{x}})(1-\rho^2_{z_jy_k\bullet\mathbf{x}})$$

$$= \sum_{k=1}^{q} (1 - \rho^2_{y_k z_j \bullet \mathbf{x}}) \text{ (using Kendall and Stuart's identity)}$$

$$= q - \sum_{k=1}^{q} \rho^2_{y_k z_j \bullet \mathbf{x}}$$

$$= q[1 - \Theta(\mathbf{x}, z_j)] \quad (j = 1,\ldots,t).$$

Hence, division by q and complementing the result to 1 yields the redundancy associated with adding z_j to the current predictor set, **x**, for j = 1,…,t. With this information, selection of z_j can be based on a criterion of maximizing conditional redundancy at each forward step.

BACK-SWITCHING

Evaluation of a redundancy after an x - z switch requires the same computational steps as previously outlined in the context of interdependence analysis. The only additional requirement here is that all correlations involving z's currently excluded from the model are excluded from the matrix to be swept. In terms of a PROC IML SWEEP of the entire correlation matrix, these elements may simply be ignored.

Example 2. Yields of 8 chlorine-homologues of dioxin and 8 chlorine-homologues of furan where measured in US EPA's Multifuel Combustor Facility under 24 different run conditions (Gullett, *et al.*, 1998). These 16 homologue yields were transformed to 15 logratios,

$$y_j = \ln[(p_j + \varepsilon) / (p_{16} + \varepsilon)] \quad j = 1,\ldots,15, \sum_{j=1}^{16} p_j = 1,$$

representing transformed homologue *composition*, using 8-Cl furan as the baseline, and ε = 0.001 as a "starter". Eleven variables were used to characterize the run conditions, namely flow rates of HCl and SO_2 in the exhaust stream; R_a and R_b and t_a and t_b, representing residence times and temperatures associated with sampling ports A and B; RDF_f, $Coal_f$, and S_f, representing feed rates of refuse derived fuel, coal, and calcium hydroxide; and Quench, a derived quench rate across the exhaust cooling section. A subset of these and the ratio $R_{S/Cl} = [SO_2]/[HCl]$ was sought as the fundamental explanatory variables for homologue compositional variation. This example is unique in that separate, univariate models for each of the responses, i.e., the logratios, would be of no interest; all 15 logratios must be known in order to predict the composition of any single homologue. Hence, the need for *multivariate multiple regression*. Maximum redundancy was used as the selection criterion for the 15-variate response, allowing 1- and 2-tuple switching.

Selected Predictors	Redundancy	Min R^2	p-value	Comments
t_b	0.058	0.001	0.589	
(t_b, HCl)	0.110	0.012	0.118	Replacement offered
R_b, $R_{S/Cl}$	0.153	0.002	0.059 0.116	2-tuple switch
R_b,$R_{S/Cl}$, R_a	0.221	0.002	0.487	
R_b,$R_{S/Cl}$,R_a, SO_2	0.312	0.010	0.745	
R_b,$R_{S/Cl}$,R_a, SO_2, S_f	0.397	0.170	0.511	
R_b,$R_{S/Cl}$,R_a, SO_2,S_f, $Coal_f$	0.452	0.263	0.267	

This example is unique in that, aside from a 2-tuple switch, the algorithm proceeded without further pause to a 6-predictor model. By contrast, use of the max-min R^2 criterion resulted in switches at each step except for 1- and 5-term models. Even then, min R^2=0.126 for the 5-term model (S_f, R_a, RDF_f, t_a, Quench), and min R^2=0.218 for the 6-term model (S_f, R_a, R_b, RDF_f, $R_{S/Cl}$, Quench), i.e., both inferior to the above results which focused on maximizing redundancy. While average R^2 showed continued improvement as (non-significant) predictors were added, min R^2 remained abysmally low, and ultimately the analysis was completed through use of generalized additive models (Dunn, 1998).

Example 3. Gullet, *et al.* (1999) reported yields of dioxin, furan, and their total from a commercial-scale, municipal waste incinerator in which *refuse derived fuel* (RDF) was cofired with either high sulfur Illinois coal or low sulfur Navy coal. Thirteen test runs were completed in a 6-day (very costly) effort. In that the RDF feeder often clogged during test runs, variables reflecting within-run variability of conditions were included in the following list of 22 potential predictors, as well as others representing combustion conditions (reflected by CO, HCl, and SO_2), and possible carryover effects:

Average feed rates (COALAV, NVCOAL, ILCOAL, RDFAV);

Average combustion conditions (COAV, HCLAV, SO2AV);

Within-run variability (RDFMIN, RDFRANG, RDFFR, RDFADJ, COMIN, CORANG, COFR, COADJ);

Exponentially-smoothed, lagged effects (COAL1, NV1, IL1, RDF1, HCL1, and S1 and S2 representing lagged SO_2 effects).

Max-min R^2 was used as the selection criterion for logarithms of the 3-variate response, allowing 1- and 2-tuple switching.

Selected Predictors	Redundancy	Min R^2	p-value	Comments
COALAV	0.515	0.368	0.011	
(COALAV, S2)	0.710	0.665	0.097	Replacement offered
RDFMIN, RDFRANG	0.755	0.721	0.022 0.011	2-tuple switch
RDFMIN, RDFRANG, RDF1	0.880	0.852	0.067	
RDFMIN, RDFRANG, RDF1, S1	0.907	0.899	0.067	
(RDFMIN, RDFRANG, RDF1, S1, HCLAV)	0.930	0.919	0.526	Replacement offered
RDFMIN, RDFRANG, S1, COFR, COAL1	0.929	0.920	0.129 0.093	2-tuple switch

By contrast, the 5 predictors selected using the maximum redundancy criterion were RDF1, S1, IL1, HCL1, and RDFRR with Θ=0.941 representing considerable gain, but min R^2=0.901 representing considerable penalty compared to the above. This example is interesting in that improvement was possible only by use of 2-tuple switching. The need for analysis of a trivariate response, rather than three separate univariate responses, is less obvious here than in the previous example, but is motivated by the need to identify factors which simultaneously affected yields of both dioxin and furan.

CONCLUSION

Stepwise predictor selection in the context of multivariate multiple regression is an obvious extension of corresponding univariate algorithms. One wonders why it is not already implemented in, say, PROC REG. Interdependence analysis seems a viable substitute for principal components/factor analysis in many applications, representing honest reduction in dimensionality without the ambiguity of identifying derived variables. Software limited to 1-tuple switches, as in PROC REG/SELECTION=MAXR, possibly reflects cycle times of contemporary hardware. This needs to be re-examined. We have been surprised at the frequency with which k(>1)-tuple switching has led to model improvement. There appears to be no consistent advantage in execution times between optimizing either maximum redundancy or max-min R^2. CPU times for PROC IML ranged from 1.35 to 10.95 seconds for the examples presented here, executing on a heavily time-shared Sun Ultra-Enterprise 5000. Source code and the data for examples presented here can be obtained from the second author (jdunn@comp.uark.edu), specifying University of Arkansas, Statistical Laboratory, Technical Report No. 28.

REFERENCES

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

Boyce, D.E., Farhi, A., Weischedel, R. (1974). *Optimal Subset Selection: Multiple Regression, Interdependence and Optimal Network Algorithms.* Lecture Notes in Economics and Mathematical Systems No. 103, Springer-Verlag.

Divekar, D.A., Dutton, R.W., and McCalla, W.J. (1977). Experimental study of Gummel-Poon model parameter correlations for bipolar junction transistors. *IEEE Journal of Solid-State Circuits* 12: 552-559.

Dunn, J.E. (1998). Case studies of vector generalized additive models in environmental health and combustion research *9th International Conference on Quantitative Methods in the Environmental Sciences (TIES98)*, Queensland, Australia, and submitted to *Environmetrics*.

Garimella, H. (1997). Model Parameter Correlations for the N21S7X35 NPN Series Transistors. M.S. thesis, Department of Electrical Engineering, University of Arkansas.

Gullett, B.K., Dunn, J.E., Bae, S., and Raghunathan, K. (1998). Mechanistic implications of polychlorinated dibenzodioxin and dibenzofuran homologue profiles from municipal waste and coal co-combustion. *Journal of Waste Management* (in press).

Gullett, B.K., Dunn, J.E., Bae, S., and Raghunathan, K. (1999). The effect of coal sulfur in reducing formation of polychlorinated dibenzo-p-dioxin and dibenzofuran during waste combustion (in preparation).

Kendall, M.G., and Stuart, A. (1979). *The Advanced Theory of Statistics*, Vol. 2, 4th Edition, page 355. Macmillan Pub. Co.

National Research Council (1996). *Massive Data Sets: Proceedings of a Workshop*. Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Lynette Duncan or James E. Dunn
University of Arkansas
SCEN 301
Fayetteville, AR 72701
Work Phone: (501) 575-3351
Fax: (501) 575-8630
Email: duncan@comp.uark.edu or jdunn@comp.uark.edu

SAS and SAS/IML software are registerd trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.