An Application of the Generalized Shrunken Least Squares Estimator on Principal Component Regression

Professor Jann-Huei Jinn
Department of Statistics
Grand Valley State University
Allendale, MI 49401
USA

Professor Chwan-Chin Song and Mr. W. B. Mou
Department of Applied Mathematics
National Cheng-Chi University
Taipei, Taiwan, R.O.C.

1. Introduction

When colinearity exists in a study of multiple linear regression, it will yield large estimated variances for the estimated coefficients in the model and it is then difficult to detect the "significant" regression coefficients. The problems caused by colinearity can be overcome somewhat by (1) deleting predictor variables that are strongly correlated (2) relating the response variable to the principal components of the predictor variables. The response variable is then regressed on these new predictor variables (3) increase sample size.

This study will propose a modified method to delete predictor variables that are strongly correlated. This method is based on Li-Chun Wang's (Generalized Shrunken Least Squares Estimator ,1990) idea and the method of deleting predictor variables by Mansfield, Webster, and Gunst (1977). We believe the modified method will reduce more estimated variances for the estimated coefficients in the model.

We plan to use real data sets to compare our modified method and the method proposed by Mansfield, Webster, and Gunst (1977).

2. Multiple Linear Regression Model and Multicollinearity

2.1 The multiple linear regression model can be written as

$$\underline{Y} = \beta_o \mathbf{1} + \mathbf{X}\underline{\beta} + \underline{\varepsilon} \qquad (2.1)$$

$$\underline{\varepsilon} \sim N_n(0, \sigma^2 I_n)$$

Where $\underline{Y}$ is a $n \times 1$ response vector

β_o is an unknown constant term

$\mathbf{1}$ is a $n \times 1$ vector with all elements 1

$\mathbf{X}$ is a $n \times k$ design matrix with all the elements has been standardized, i.e., $\sum_i X_{ij} = 0$ and $\sum_i X_{ij}^2 = 1$.

$\underline{\beta}$ is a $k \times 1$ parameter vector

$\underline{\varepsilon}$ is a $n \times 1$ random error vector

I_n is a $n \times n$ identity matrix of order n

A very desirable condition in a set of regression data is to have "orthogonality". The condition of orthogonality of an experiment design occurs when one truly does have the capability to control the regressor variables. Let $\mathbf{X}=\lfloor \underline{x_1}, \underline{x_2}, ..., \underline{x_k} \rfloor$ where $\underline{x_j}$ is a $n \times 1$ vector of n observations for the jth predictor.

Multicollinearity simply occurs when there are near linear dependence among the $\underline{x_j}$, the columns of **X**. That is, there is a set of constants c_j (not all zero) for which

$$\sum_{j=1}^{k} c_j \underline{x_j} \cong 0 \qquad (2.2)$$

We write $\cong$ since if the right hand side is identically zero, the linear dependencies are exact and thus (1/**X'X**) does not exist. Near dependencies, of course, may exist in real data and produce the effect that commonly call multicollinearity. One should keep in mind that a regression coefficient is a rate of change or partial derivate of the response with respect to a regressor variable. When the x-data are conditioned in such a way that the regressor are moving with one another, the least squares procedure never is allowed exposure to the data structure that it truly needs to produce a clear estimate of this rate of change.

Let $\underline{\hat{\beta}}$ =X'Y/(**X'X**) represents the vector of least squares estimates of $\underline{\beta}$ in the model of (2.1). The covariance matrix of $\underline{\hat{\beta}}$ is V($\underline{\hat{\beta}}$)= σ^2 /(**X'X**). Suppose we consider the **X'X** matrix (correlation form). We know that there exists an orthogonal matrix (see Grabill (1976))

$$\mathbf{V}=\lfloor \underline{v_1}, \underline{v_2}, ..., \underline{v_k} \rfloor \text{ such that } \mathbf{V'(X'X)V}=\text{diag}(\lambda_1, \lambda_2, ..., \lambda_k). \qquad (2.3)$$

The λ_i are the *eigenvalues* of the correlation matrix. The operation given by (2.3) is called the *eigenvalue decomposition* of **X'X.** The columns of **V** are normalized eigenvectors associated with the eigenvalues of **X'X.** For our purposes here, we need to denote the ith element of the vector $\underline{v_j}$ by v_{ij}. Now if multicollinearity is present, at least one $\lambda_i \cong 0$. Thus we can write, for at least one value of j, $\underline{v_j}'$ (**X'X**) $\underline{v_j} \cong 0$ which implies that for at least one eigenvector $\underline{v_j}$, $\sum_{l=1}^{k} v_{lj} \underline{x_l} \cong 0$. Thus the number of small eigenvalues of the correlation matrix relate to the number of multicollinearity according to the definition in (2.2), and the "weights", the c_j in (2.2), are the individual elements in the associated eigenvectors.

Define **D**= (1/**X'X**), the diagonal elements of this matrix can be expressed as $d_{jj} = \frac{1}{1-R_j^2}$, $j = 1,2,...,k$ where R_j^2 is the coefficient of multiple determination of the regression produced by regression the variable $\underline{x_j}$ against the other regressor variables, the $\underline{x_i}$ ($j \neq i$). Thus the higher the multiple correlation in the regression, the lower the precision in the estimate of the coefficient $\hat{\beta}_j$. That is, when R_j^2 approaches to 1, the variance of individual least squares estimator $\hat{\beta}_j$, Var($\hat{\beta}_j$)= $d_{jj}\ \sigma^2 = \sigma^2 (\frac{1}{1-R_j^2})$,will become very large.

Define the squared distance between the estimated vector $\underline{\hat{\beta}}$ and the true parameter vector $\underline{\beta}$ as L_1^2=($\underline{\hat{\beta}}$ - $\underline{\beta}$)'($\underline{\hat{\beta}}$ - $\underline{\beta}$). The other approach in illustrating the effect of multicollinearity on the coefficients is the consideration of the expected squared distance between the estimated vector $\underline{\hat{\beta}}$ and the true parameter vector $\underline{\beta}$. i.e.,

$$E(L_1^2)=E[(\underline{\hat{\beta}} - \underline{\beta})'(\underline{\hat{\beta}} - \underline{\beta})] = \sigma^2 \text{tr}(1/\mathbf{X'X}) = \sigma^2 \sum_{j=1}^{k} \frac{1}{\lambda_j} \qquad (2.4)$$

where λ_j >0 are the *eigenvalues* of the correlation matrix, $j = 1,2,...,k$.

Thus for an ill-conditioned or near singular **X'X,** at least one of the eigenvalues will be small and thus E[($\underline{\hat{\beta}}$ - $\underline{\beta}$)'($\underline{\hat{\beta}}$ - $\underline{\beta}$)] will be large. For the ideal case (orthogonality), $\sum_{j=1}^{k} \frac{1}{\lambda_j}$ =k. It becomes clear, then, from equation (2.4), that since

$$E[(\underline{\hat{\beta}} - \underline{\beta})'(\underline{\hat{\beta}} - \underline{\beta})] = E(\underline{\hat{\beta}}'\underline{\hat{\beta}}) - (\underline{\beta}'\underline{\beta})$$

then
$$E(\underline{\hat{\beta}}'\underline{\hat{\beta}}) = (\underline{\beta}'\underline{\beta}) + \sigma^2 \sum_{j=1}^{k} \frac{1}{\lambda_j} \qquad (2.5)$$

Equation (2.5) underscores the tendency for multicollinearity to produce a vector of regression coefficients that is too long, i.e., coefficients that have the tendency to be *too large* in magnitude. If any of the λ_j are small, obviously ($\underline{\hat{\beta}}'\underline{\hat{\beta}}$) is heavily biased upward for ($\underline{\beta}'\underline{\beta}$), and hence one would expect coefficients that are large. This is true in spite of the fact that the $\hat{\beta}_j$'s themselves are unbiased. For example, an eigenvalue of 0.0005 is not at all uncommon in highly collinear situations. Clearly, from (2.5), for this situation, $\sum_{j=1}^{k} \hat{\beta}_j^2$ is heavily biased, and the result is a tendency for some of the coefficients to be overestimated in magnitude.

2.2 Multicollinearity Diagnostics

The following represent formal multicollinearity diagnostic tools.

a) Simple Correlations Among the Regressor Variables

The analyst normally has access to the correlation matrix of the regressor variables, i.e., **X'X.** These numbers , of course, indicate pairwise type correlations. However, we should clarify that multicollinearity quite often involves associations among multiple regressor variables. As a result, the simple correlations themselves do not always underscore the extent of the problem. There are no definite guideline values on the simple correlations and, while they should be observed so that the analyst can see which one-on-one associations exist, they do not always indicate the actual nature or the extent of the multicollinearity.

b) Variance Inflation Factors

Define VIF= $\frac{1}{1-R_j^2}$ as the variance inflation factor. The VIFs represent the inflation that each regression coefficient experiences above ideal, i.e., above what would be experienced if the correlation matrix were an identity matrix. It is easy to see that it involves the notion of multiple association. If R_j^2 is near unity, $(VIF)_j$ will be quite large. This will occur if the ith regressor variable has a strong linear association with the remaining regressors. The VIFs represent a considerably more productive approach for detection than do the simple correlation values. They supply the user with an indication of which coefficients are adversely affected and to what extent. It is generally believed that if any VIF greater than 10, there is reason for at least some concern; then one should consider variable deletion or an alternative to least squares estimation to combat the problem.

c) Condition Number of the Correlation Matrix

We know that eigenvalues and eigenvectors of the correlation matrix, **X'X,** play an important role in the multicollinearity that exists in a set of regression data. Indeed, the nearness to zero of the smallest eigenvalue is a measure of the strength of a linear dependency, while the elements of the associated normalized eigenvector display the *weights* on the corresponding regressor variables in the multicollinearity. Of course, the eigenvalues would all be 1 if the variables define an orthogonal system so this provides a norm for the analyst. In addition, the *spectrum* of eigenvalues produces another diagnostic. Multicollinearity can be measured in terms of the ratio of the largest to the smallest eigenvalues.

Define the condition number of the correlation matrix as

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$$

Large values of κ are an indication of serious multicollinearity. An excessively large κ is evidence that the regression coefficients are unstable, i.e., subject to major changes with small perturbations in the regression data. Numerical rules of thumb says κ <100 i indicates a weak multicollinearity, 100< κ <1,000 indicates a moderate multicollinearity, if κ >1,000 one should be concerned about the effect of multicollinearity. The condition number κ is more reliable for diagnosing the impact of a dependency than the eigenvalue λ_j itself.

3. Alternatives to Least Squares When Multicollinearity Exists

There are many estimation procedures designed to combat multicollinearity, procedures that were developed to eliminate model instability and to reduce the variances of the regression coefficients.

At the point in which the analyst has determined, by the use of the diagnostics, that Multicollinearity is a problem, often a substantial benefit may be derived from an attempt to eliminate much of the multicollinearity *without* resorting to alternatives to least squares. The very presence of multicollinearity in the diagnostics suggests that, in the case of k regressor variables, the actual model-building exercise should involve fewer than k variables. In other words, there is not sufficient information in the regressor data to warrant modeling k regressors. As a result, the analyst often can eliminate or certainly reduce the effect of multicollinearity by removing one or more regressors.

3.1 Ridge Regression

Ridge regression is one of the most popular, though controversial, estimation procedures for combating multicollinearity. The procedures related to Ridge regression fall into the category of *biased estimation techniques*. They are based on this notion: though ordinary least squares gives unbiased estimates and indeed enjoy the minimum variance of all linear unbiased estimators, there is no upper bound on the variance of the estimators and the presence of multicollinearity may produce very large variances. As a result, one can visualize that, under the condition of multicollinearity, a huge price is paid for the unbiasedness property that one achieves by using ordinary least squares. Biased estimation is used to attain a substantial reduction in variance with an accompanied increase in stability of the regression coefficients.

The ridge regression estimator of the coefficient $\underline{\beta}$ is found by solving for $\underline{\hat{\beta}_R}$ in the system of equation

$$(\mathbf{X'X} + \phi\, I_n)\,\underline{\hat{\beta}_R} = \mathbf{X'}\,\underline{Y} \qquad (3.1)$$

where $\phi \geq 0$ is often referred to as a *shrinkage parameter*. The solution, of course, is given by

$$\underline{\hat{\beta}_R} = [1/(\mathbf{X'X} + \phi\, I_n)]\mathbf{X'}\,\underline{Y} \qquad (3.2)$$

There are various procedures for choosing the shrinkage parameter ϕ. A fairly simple study of the properties of the ridge estimator in (3.2) reveals the role of ϕ in moderating the variance of the estimators. Perhaps the most dramatic illustration of the impact of small eigenvalues on the variances of the least squares coefficients is the expression for $\sum_j (Var\hat{\beta}_j)/\sigma^2$ given in equation (2.5). In the case of the ridge regression estimator, the equivalent property is given by

$$\sum_{j=1}^{k} \frac{Var\hat{\beta}_{j,R}}{\sigma^2} = \sum_{j=1}^{k} \frac{\lambda_j}{(\lambda_j + \phi)^2} \qquad (3.3)$$

For example, in the case of k=3 regressor variables with λ_1=2.985, λ_2=0.01, and λ_3=0.005, least squares estimation gives

$$\frac{\sum_{j=1}^{3} Var\hat{\beta}_j}{\sigma^2} = \sum_{j=1}^{3} \frac{1}{\lambda_j} = 0.3350+100+200=300.3350$$

If ridge regression with say ϕ=0.10 is used, the sum of the variances is given by

$$\sum_{j=1}^{3} \frac{\lambda_j}{(\lambda_j + \phi)^2} \cong 2.3$$

It is clear when multicollinearity is severe, i.e., when there is at least one near zero eigenvalue, much improvement in variance, and thus coefficient stability, can be experienced. Equation (3.3) emphasizes that the ϕ in ridge regression moderates the damaging impact of the small eigenvalues that result from the collinearity.
The bias that results for a selection of ϕ>0 is best quantified by observing an expression for $\sum_{j=1}^{k} (Bias\hat{\beta}_{j,R})^2 = \sum_{j=1}^{k} \left[E(\hat{\beta}_{j,R}) - \beta_j\right]^2$, the sum of the squared biases of the regression coefficients. This expression is given by (see Hoerl and Kennard (1970(a))

$$\sum_{j=1}^{k} \left[E(\hat{\beta}_{j,R}) - \beta_j\right]^2 = \phi^2\, \underline{\beta}\,\mathbf{'}[\mathbf{X'X} + \phi\, I_n\,]^{-2}\, \underline{\beta} \qquad (3.4)$$

Thus we can expect that the procedure of ridge regression would be successful if a ϕ is chosen so that the variance reduction is greater than the bias term given in (3.4). There is no assurance that this can be done because the analyst will never know what the bias is. The choice of ϕ belongs to the analyst, of course, and a parameter value should be chosen where results show strong evidence that improvements in the estimates are being experienced.

3.2 Principal Component Regression

Principal components regression represents another biased estimation technique for combating multicollinearity. We perform least squares estimation on a set of variables called the *principal components* of the correlation matrix. Based on the nature of the analysis, we delete a certain number of the principal components to effect a substantial reduction in variance. The method varies somewhat in philosophy from ridge regression but, like ridge regression, gives *biased* estimates; when used successfully, this method results in estimation and prediction that is superior to OLS (ordinary least squares). Principal components are orthogonal to each other, so that it becomes quite easy to attribute a specific amount of variance to each.

Consider the matrix of normalized eigenvectors associated with the eigenvalues $(\lambda_1, \lambda_2, ..., \lambda_k)$ of **X'X** (correlation form). Let **V**= $\lfloor \underline{v_1}, \underline{v_2}, ..., \underline{v_k} \rfloor$ be a k× k orthogonal matrix where the jth column vector $\underline{v_j}$ of **V** is a normalized eigenvector associated with the jth eigenvalue λ_j of **X'X.** We know that **VV'=V'V=I** since **V** is an orthogonal matrix. Hence we can write the original regression model in the form

$$\underline{Y} = \beta_o \mathbf{1} + \mathbf{XVV'}\,\underline{\beta} + \underline{\varepsilon} \qquad (3.5)$$

$$\underline{Y} = \beta_o \mathbf{1} + \mathbf{W}\,\underline{\gamma} + \underline{\varepsilon} \qquad (3.6)$$

where **W** =**XV** and $\underline{\gamma}$ **=V'** $\underline{\beta}$ **.** **W** is an n× k matrix and $\underline{\gamma}$ is a k× 1 vector of new coefficients $\gamma_1, \gamma_2, ..., \gamma_k$. We can visualize the columns of **W** (typical element w_{ij} as representing readings on *k new variables, the principal components*. It is easy to see that the components are orthogonal to each other. We have

$$\mathbf{W'W=(XV)'(XV)=V'X'XV} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_k) = \Lambda \qquad (3.7)$$

So, if regression is performed on the w's via the model in (3.6), the variances of coefficients (the diagonal elements of $(\mathbf{W'W})^{-1}$ apart from σ^2) are the reciprocals of eigenvalues. That is,

$$\text{Var}(\hat{\gamma}_j) = \frac{\sigma^2}{\lambda_j} \qquad j = 1,2,...,k$$

Note that the $\hat{\gamma}_j$'s are the least squares estimators ($\underline{\hat{\gamma}} = \Lambda^{-1}\mathbf{W'}\,\underline{Y}$).
The least squares prediction equation then becomes

$$\underline{\hat{Y}} = \underline{\overline{Y}} + \sum_{j=1}^{k} u_j \hat{\gamma}_j \qquad (3.8)$$

where u_j = **z'** $\underline{v_j}$ is the jth principal component value for the point at which prediction is desired, **z**. If there are s principal components associated with the s largest latent roots were retained, then $\underline{Y^*} = \underline{\bar{Y}} + \sum_{j=1}^{s} u_j \hat{\gamma}_j$ (3.9)

Based on (3.9) the residual sum of squares for $\underline{Y^*}$ is

$$SSE^* = \sum_{i=1}^{n} (\underline{Y_i} - \underline{Y_i^*})^2 = \sum_{i=1}^{n} (\underline{Y_i} - \underline{\bar{Y}})^2 - \sum_{j=1}^{s} \lambda_j \hat{\gamma}_j^2$$

3.3 Generalized Shrunken Least Squares Estimator

Li-Chun Wang (1990) proposed Generalized Shrunken Least Squares Estimator (GSLSE),

$$\underline{\hat{\beta}_{GS}} = \mathbf{VAV'}\,\underline{\hat{\beta}}\ , \qquad (3.10)$$

to estimate $\underline{\beta}$ in the model $\underline{Y} = \beta_o \mathbf{1} + \mathbf{X}\underline{\beta} + \underline{\varepsilon}$ **.** Where **A**=diag$(a_1, a_2, ..., a_k)$, $0 \le a_j \le 1$, $j = 1,2,...,k$, and **V**=$\lfloor \underline{v_1}, \underline{v_2}, ..., \underline{v_k} \rfloor$ is an orthogonal matrix such that **V' X'XV= V'(X'X)V**=diag$(\lambda_1, \lambda_2, ..., \lambda_k)$= Λ. For convenience, λ_j's satisfy $\lambda_1 \ge \lambda_2 \ge ... \ge \lambda_k > 0$. Wrung-Shen Yeh (1991) derived the leverage formula and some properties for GSLSE. He also proved that when there is severe multicollinearity, GSLSE provides better estimates than least squares estimates, $\underline{\hat{\beta}}$.

4. Variable Selection Methods

4.1 MWG Variable Selection Technique

Mansfield, E. R., Webster, J. T., and Gunst, R. F. (MWG,1977) presented an analytic technique for deleting predictor variables from a linear regression model when principal components of X'X are removed to adjust for multicollinearities in the data.

Consider model (3.6) and partition **V**=$[V_s : V_{k-s}]$, where $V_s = [V_1, V_2, ..., V_s]$ contains the latent vectors corresponding to the s largest latent roots. Consequently,

$$\mathrm{W}=[W_s : W_{k-s}]=\mathrm{X}[V_s : V_{k-s}].$$

Then model (3.6) can be rewritten as

$$\underline{Y} = \beta_o \mathbf{1} + [W_s : W_{k-s}] \begin{bmatrix} \gamma_s \\ \gamma_{k-s} \end{bmatrix} + \underline{\varepsilon}$$

Let Λ_s=diag$(\lambda_1, \lambda_2, ..., \lambda_s)$. Then if $\underline{\gamma_s} = (\gamma_1, \gamma_2, ..., \gamma_s)'$, the $\hat{\gamma}_j$ used in (3.9) can be obtained as $\underline{\hat{\gamma}_s} = (\hat{\gamma}_1, \hat{\gamma}_2, ..., \hat{\gamma}_s)' = \Lambda_s^{-1} W_s' \underline{Y}$. (4.1)

Considering deleting r independent variables from the predictor (3.9). Since the ordering of the variables in **X** is arbitrary, suppose the ones to be deleted are the last r variables. Now partition $V_s = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$ where V_1 is an (k-r)$\times$ s matrix and V_2 is an r$\times$ s matrix whose rows correspond to the r variables to be deleted from the predictor, i.e., the last rows of V_s. In order to obtain a reduced predictor with the r variables deleted, MWG considered an alternative estimator of $\underline{\gamma_s}$:

$$\tilde{\gamma}_s = A_s \, \Lambda_s^{-1} \, W_s' \, \underline{Y} = A_s \, \underline{\hat{\gamma}_s} \, . \qquad (4.2)$$

where A_s =diag$(a_1, a_2, ..., a_s)$. Note that if $A_s = I_s$, $\tilde{\gamma}_s = \underline{\hat{\gamma}_s}$.

The estimator of $\underline{\beta}$ corresponding to (4.2) is $\underline{\tilde{\beta}} = V_s \tilde{\gamma}_s = \begin{pmatrix} V_1 A_s \hat{\gamma}_s \\ V_2 A_s \hat{\gamma}_s \end{pmatrix} = \begin{pmatrix} \tilde{\beta}_{k-r} \\ \tilde{\beta}_r \end{pmatrix}$, (4.3)

where the last r rows in (4.3) correspond to the independent variables which are to be deleted. MWG showed that a matrix A_s can be found which minimizes the residual sum of squares subject to $V_2 \; A_s \; \tilde{\gamma}_s$ =**0**, i.e., $\tilde{\beta}_r$ =**0**, which removes the last r variables from the resulting predictor.
Use of the principal component estimator with r variables deleted results in a prediction equation

$$\underline{\tilde{Y}} = \underline{\bar{Y}} + \sum_{j=1}^{s} u_j \tilde{\gamma}_j = \underline{\bar{Y}} + \sum_{j=1}^{k-r} z_j \tilde{\beta}_j \, , \qquad (4.4)$$

where z_j is a value of the jth independent variable from the original model (2.1). The residual sum of squares for (4.4) can be written as

$$\underline{SSE} = \sum_{i=1}^{n} (\underline{Y_i} - \underline{\tilde{Y}_i})^2 = SSE^* + \sum_{j=1}^{s} (1 - a_j)^2 \lambda_j \hat{\gamma}_j^2 = \; SSE^* + u_r \qquad (4.5)$$

where u_r is the increase in the residual sum of squares from (3.9).

Since the original ordering of the vectors of the **X** matrix is arbitrary let the r $\underline{X_j}$'s to be deleted be the r right-hand columns of **X**. The problem is then to minimize the increase in residual sum of squares, $u_r = \sum_{j=1}^{s} (1 - a_j)^2 \lambda_j \hat{\gamma}_j^2$ subject to the restrictions $\tilde{\beta}_r = V_2 \; A_s \; \tilde{\gamma}_s$ =**0.**
MWG(1977) proved the elements of the diagonal matrix A_s is

$$a_j = 1 - \{ v_{2j}' [V_2 \; \Lambda_s^{-1} \; V_2' \,]^{-1} \, V_2 \; \underline{\hat{\gamma}_s} \, / \lambda_j \hat{\gamma}_j \} \qquad j = 1,2,...,s \qquad (4.6)$$

$$\underline{\hat{\beta}_{k-r}} = V_1 [\mathbf{1} - \Lambda_s^{-1} \, V_2' (V_2 \; \Lambda_s^{-1} \; V_2' \,)^{-1} \, V_2 \,] \underline{\hat{\gamma}_s} \qquad (4.7)$$

and

$$u_r = \sum_{j=1}^{s} (1 - a_j)^2 \lambda_j \hat{\gamma}_j^2 = \underline{\hat{\gamma}_s}' \, V_2' (V_2 \; \Lambda_s^{-1} \; V_2' \,)^{-1} \, V_2 \; \underline{\hat{\gamma}_s} \qquad (4.8)$$

where $V_2 = (v_{21}, v_{22}, ..., v_{2s})$.

MWG (1977) proposed the following steps to delete independent variables:

Step1: The value of u_1 would be computed for each of the k variables in (3.9) by using (4.8). denote the smallest of these k u_1's by u_{1m}. Using MSE= $\frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-k-1}$ (the least squares mean squared error) from the full model, F=u_{1m}/MSE can be used to determine whether this variable should be deleted. If u_{1m} is sufficiently small (the test is not significant) delete the corresponding $\underline{X_j}$ from the model and proceed to step2.

Step2: The second step deletes the variable removed previously and each of the remaining $\underline{X_j}$ (one at a time), calculating u_2 for each of the k-1 pairs of variables. Denote the smallest of the u_2 values by u_{2m}. Use F=(u_{2m}-u_{1m})/MSE to determine whether these variables should be deleted. If u_{2m}-u_{1m} is sufficiently small, both the variables involved are deleted; otherwise, only the variable deleted in step 1 is removed and the process is terminated.

Step3: At this step the smallest of k-2 u_3 values is determined by calculating (4.8) for the two variables removed in step 2 and each of the ones remaining. Denote the smallest of the u_3 values by u_{3m}. Use F=(u_{3m}- u_{2m})/MSE to determine whether these variables should be deleted. If u_{3m}- u_{2m} is sufficiently small, all three variables involved are deleted and the procedure is continued; otherwise, the process is discontinued with only the two variables deleted at step 2 removed. The k-l variables remaining when the elimination stops (i.e., l variables are deleted) are considered important independent variables yielding information on the dependent variable, Y.

The j+1 steps use F=($u_{(j+1)m} - u_{jm}$)/MSE , j=0,1,2,…,l as the test statistic value where u_{0m}=0 and l is the number of variables deleted.

An alternative backward elimination-type technique for deletion of variables is to proceed as in the first step above, but reevaluate the model at each step. In other words, at step 2 examine the latent roots and vectors of the (k-1)× (k-1) reduced **X'X** matrix, remove components corresponding to small latent roots, and then calculate a new u_{1m} from the k-1 u_1 values. Using MSE from the full model, F=u_{1m}/MSE can again be used to determine whether this second variable should be deleted.

The authors (MWG) found that this latter backward elimination has performed more satisfactory than the above method on several data sets they examined.

4.2 The Relationship Between MWG Estimator and Generalized Shrunken Least Squares Estimator

The *MWG estimator* is defined as $\underline{\tilde{\beta}} = V_s \tilde{\gamma}_s = \begin{pmatrix} V_1 A_s \hat{\gamma}_s \\ V_2 A_s \hat{\gamma}_s \end{pmatrix} = \begin{pmatrix} \tilde{\beta}_{k-r} \\ \tilde{\beta}_r \end{pmatrix}$ (4.9)

The *generalized shrunken least squares estimator* is defined as

$$\underline{\hat{\beta}_{GS}} = \mathbf{VAV'}\,\underline{\hat{\beta}} \quad (4.10)$$

where $\mathbf{A}=\mathrm{diag}(a_1, a_2, ..., a_k)$, $0 \le a_j \le 1$, $j = 1,2,...,k$, and $\mathbf{V}=\lfloor \underline{v_1}, \underline{v_2}, ..., \underline{v_k} \rfloor$ is an orthogonal matrix such that $\mathbf{V'\,X'XV = V'(X'X)V}=\mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_k)=\Lambda$.
The jth column vector $\underline{v_j}$ of **V** is a normalized eigenvector associated with the jth eigenvalue λ_j of **X'X.** For convenience let λ_j s satisfy $\lambda_1 \ge \lambda_2 \ge ... \ge \lambda_k > 0$.

Define the *most generalized ridge estimator* of $\underline{\beta}$ in the multiple linear regression model $\underline{Y} = \beta_o \mathbf{1} + \mathbf{X}\underline{\beta} + \underline{\varepsilon}$ as $\underline{\hat{\beta}_{MGR}} = \mathbf{VAV'}\,\underline{\hat{\beta}}$ (4.11)

where $\underline{\hat{\beta}}$ is the least squares estimator of $\underline{\beta}$, $\mathbf{A}=\mathrm{diag}(a_1, a_2, ..., a_k)$ is a diagonal matrix with $-\infty < a_j < \infty$, $j = 1,2,...,k$.
If we change $-\infty < a_j < \infty$ in the diagonal matrix **A** of (4.11) to $0 \le a_j \le 1$ then $\underline{\hat{\beta}_{MGR}}$ is the $\underline{\hat{\beta}_{GS}}$ of Li-Chun Wang's (Generalized Shrunken Least Squares Estimator ,1990). In the *MWG* method they partitioned $\mathbf{V}=[V_s : V_{k-s}]$, where $V_s = [V_1, V_2, ..., V_s]$ contains the latent vectors corresponding to the s largest latent roots. If we let A_{k-s} =0 in the diagonal matrix $\mathbf{A}=\begin{bmatrix} A_s & 0 \\ 0 & A_{k-s} \end{bmatrix}$ then we can obtain $\underline{\tilde{\beta}} = V_s \tilde{\gamma}_s = \begin{pmatrix} V_1 A_s \hat{\gamma}_s \\ V_2 A_s \hat{\gamma}_s \end{pmatrix} = \begin{pmatrix} \tilde{\beta}_{k-r} \\ \tilde{\beta}_r \end{pmatrix}$.
Therefore, $\underline{\tilde{\beta}}$ and $\underline{\hat{\beta}_{GS}}$ are special cases of $\underline{\hat{\beta}_{MGR}}$. If we want to improve the precision of $\underline{\tilde{\beta}}$ then we can try to confine the diagonal elements a_j of A_s in [0,1] then in this case, $\underline{\tilde{\beta}}$ becomes a special case of $\underline{\hat{\beta}_{GS}}$.

This study is to apply the constraint of $0 \le a_j \le 1$, $j = 1,2,...,k$ in the generalized shrunken least squares estimator to the method of deleting regressors proposed by Mansfield et al. (1977). We believe that the application of the constraint to the MWG estimator will get a better result of variable selection procedure.

4.3 Some Properties of MWG Estimator

Since $\tilde{\gamma}_s = A_s\, \Lambda_s^{-1}\, W_s{}'\, \underline{Y} = A_s\, \underline{\hat{\gamma}_s} = [\mathbf{1} - \Lambda_s^{-1} V_2{}' (V_2\, \Lambda_s^{-1}\, V_2{}')^{-1} V_2]\, \underline{\hat{\gamma}_s}$ (4.12),

the expected value of $\tilde{\gamma}_s$ is $E(\tilde{\gamma}_s) = [\mathbf{1} - \Lambda_s^{-1} V_2{}' (V_2\, \Lambda_s^{-1}\, V_2{}')^{-1} V_2]\, \underline{\gamma_s}$ (4.13)

and variance-covariance $\mathrm{Var}(\tilde{\gamma}_s)=\sigma^2[\Lambda_s^{-1}-\Lambda_s^{-1}V_2'(V_2\Lambda_s^{-1}V_2')^{-1}V_2\Lambda_s^{-1}]$. (4.14)
The bias of $\tilde{\gamma}_s$ is $\mathrm{E}(\tilde{\gamma}_s)-\underline{\gamma_s}=-[\Lambda_s^{-1}V_2'(V_2\Lambda_s^{-1}V_2')^{-1}V_2]\underline{\gamma_s}$

Since variance-covariance matrix of $\underline{\hat{\gamma}_s}$ is $\mathrm{Var}(\underline{\hat{\gamma}_s})=\sigma^2\Lambda_s^{-1}$ and $\Lambda_s^{-1}V_2'(V_2\Lambda_s^{-1}V_2')^{-1}V_2\Lambda_s^{-1}$ is a positive definite matrix, it implies $\mathrm{V}(\tilde{\gamma}_s)\le\mathrm{V}(\underline{\hat{\gamma}_s})$. That is, MWG estimator is better than the least squares estimator. The derivations of (4.13) and (4.14) are given in Appendix A.

In general, the least squares estimator has the following two properties: (1) The sum of residuals is equal to zero. (2) The residual vector is orthogonal to the prediction vector. We will show that MWG's estimator, $\underline{\tilde{\beta}}$, satisfies both properties. The residual vector of MWG's estimator is $\tilde{e}=\underline{Y}-\underline{\bar{Y}}\,\mathbf{1}-\mathbf{X}\underline{\tilde{\beta}}$. Since $\mathbf{X}$ is a $n\times k$ design matrix with all the elements has been standardized, i.e.,

$$\sum_i X_{ij}=0 \text{ and } \sum_i X_{ij}^2=1.$$

It is easy to prove that $\mathbf{1}'\tilde{e}=\mathbf{1}'(\underline{Y}-\underline{\bar{Y}}\,\mathbf{1}-\mathbf{X}\underline{\tilde{\beta}})=\mathbf{0}$.

If we want to show that MWG's estimator has property (2) then the $\underline{\tilde{\beta}}$ in (4.9) must satisfies $\underline{\tilde{\beta}_r}=V_2A_s\underline{\hat{\gamma}_s}=\mathbf{0}$, and $\underline{\hat{\gamma}_s}'(A_s\Lambda_s-A_s\Lambda_sA_s)\underline{\hat{\gamma}_s}=\mathbf{0}$ (4.15)
(see Appendix B for the derivation of (4.15)). The problem is then to minimize

$$u_r=\sum_{j=1}^{s}(1-a_j)^2\lambda_j\hat{\gamma}_j^2$$

such that $V_2A_s\underline{\hat{\gamma}_s}=\mathbf{0}$ and $\underline{\hat{\gamma}_s}'(A_s\Lambda_s-A_s\Lambda_sA_s)\underline{\hat{\gamma}_s}=\mathbf{0}$. (4.16)
The solution of (4.16) is surprisingly the same as (4.6). This proves MWG estimator has property (2) (see Appendix C for the proof).

When there is only one independent variable deleted, i.e. r=1, we can show that a_j in (4.6) satisfies (4.16). First, $\underline{\hat{\gamma}_s}'(A_s\Lambda_s-A_s\Lambda_sA_s)\underline{\hat{\gamma}_s}$ can be rewritten as

$\sum_{j=1}^{s}a_j(1-a_j)\lambda_j\hat{\gamma}_j^2$ and a_j in (4.6) can be rewritten as $a_j=1-V_{kj}\left(\sum_{j=1}^{s}V_{kj}^2/\lambda_j\right)^{-1}\left(\frac{\sum_{j=1}^{s}V_{kj}\hat{\gamma}_j}{\lambda_j\hat{\gamma}_j}\right)$

where $V_k'=(V_{k1},V_{k2},...,V_{ks})$ is the last row of V_s. Therefore, $\sum_{j=1}^{s}a_j(1-a_j)\lambda_j\hat{\gamma}_j^2$

$$=\sum_{j=1}^{s}\lambda_j\left[1-V_{kj}\left(\sum_{j=1}^{s}V_{kj}^2/\lambda_j\right)^{-1}\left(\frac{\sum_{j=1}^{s}V_{kj}\hat{\gamma}_j}{\lambda_j\hat{\gamma}_j}\right)\right]\left[V_{kj}\left(\sum_{j=1}^{s}V_{kj}^2/\lambda_j\right)^{-1}\left(\frac{\sum_{j=1}^{s}V_{kj}\hat{\gamma}_j}{\lambda_j\hat{\gamma}_j}\right)\right]\hat{\gamma}_j^2$$

$$= \sum_{j=1}^{s} [V_{kj} (\sum_{j=1}^{s} V_{kj}^2 / \lambda_j)^{-1} (\sum_{j=1}^{s} V_{kj} \hat{\gamma}_j)] \hat{\gamma}_j - \sum_{j=1}^{s} [V_{kj}^2 (\sum_{j=1}^{s} V_{kj}^2 / \lambda_j)^{-2} (\sum_{j=1}^{s} V_{kj} \hat{\gamma}_j)^2] / \lambda_j$$

$$= (\sum_{j=1}^{s} V_{kj}^2 / \lambda_j)^{-1} (\sum_{j=1}^{s} V_{kj} \hat{\gamma}_j)^2 - (\sum_{j=1}^{s} V_{kj}^2 / \lambda_j)^{-2} (\sum_{j=1}^{s} V_{kj} \hat{\gamma}_j)^2 (\sum_{j=1}^{s} V_{kj}^2 / \lambda_j)$$

$$= (\sum_{j=1}^{s} V_{kj}^2 / \lambda_j)^{-1} (\sum_{j=1}^{s} V_{kj} \hat{\gamma}_j)^2 - (\sum_{j=1}^{s} V_{kj}^2 / \lambda_j)^{-1} (\sum_{j=1}^{s} V_{kj} \hat{\gamma}_j)^2 = 0$$

4.4 MWG Variable Selection Technique With Restrictions

In section 4.1., equation (4.2) $\tilde{\gamma}_s = A_s \Lambda_s^{-1} W_s' \underline{Y} = A_s \underline{\hat{\gamma}_s}$ or equation (4.3) $\underline{\tilde{\beta}} = V_s \tilde{\gamma}_s = V_s A_s \underline{\hat{\gamma}_s}$, there were no restrictions to the diagonal elements a_j of matrix A_s, therefore some variances could become vary large. In section 4.2 we found $\underline{\tilde{\beta}}$ and $\underline{\hat{\beta}_{GS}}$ are special cases of $\underline{\hat{\beta}_{MGR}}$. If we want to improve the precision of $\underline{\tilde{\beta}}$ then we can try to confine the diagonal elements a_j of A_s in [0,1] then in this case, $\underline{\tilde{\beta}}$ becomes a special case of $\underline{\hat{\beta}_{GS}}$. If we apply the constraint of $0 \le a_j \le 1,\ j = 1,2,...,k$ in the generalized shrunken least squares estimator to the method of deleting regressors proposed by Mansfield et al. (1977), what will be the effect to the MWG's variable selection procedure? We believe that the application of the constraint to the MWG estimator will get a better result of this variable selection technique.

In equation (4.11), $\underline{\hat{\beta}_{MGR}}$ = **VAV'** $\underline{\hat{\beta}}$, **A**=diag$(a_1, a_2, ..., a_k)$ is a diagonal matrix with $-\infty < a_j < \infty,\ j = 1,2,...,k$. We now change $-\infty < a_j < \infty$ in the diagonal matrix **A** to $0 \le a_j \le 1$. The MWG variable selection steps are similar to those steps in section 4.1.

The minimization of (4.15) can be rewritten as

$$\text{Min}\ \ u_r = \sum_{j=1}^{s} (1 - a_j)^2 \lambda_j \hat{\gamma}_j^2 \qquad (4.17)$$

$$\sum_{j=1}^{s} a_j V_{tj} \hat{\gamma}_j = 0,\ \text{t=k-r+1, k-r+2,…,k} \qquad (4.18)$$

$$\text{and}\ 0 \le a_j \le 1, \quad j = 1,2,...,s \qquad (4.19)$$

where V_{tj} are elements of $V_2 = (v_{21}, v_{22}, ..., v_{2s})$ (an r× s matrix whose rows correspond to the r variables to be deleted from the predictor, i.e., the last rows of V_s).

To find the minimum value of (4.17), we propose the following two methods: (1) Ignore the constraint of (4.18) and use Lagrange multiplier to find the values of a_j (see equation (4.6)). Then use Rao-Ghangurde method to adjust a_j so that (4.18) is satisfied. If all the a_j s are in $[0,1]$ then substitute them into (4.16) to find u_r; if some a_j are not fall within $[0,1]$, the Rao-Ghangurde adjustment method is used as follows: If $a_j > 1$ then use $a_j = 1$; if $a_j < 0$ then use $a_j = 0$. Substituting these adjusted a_j s (along with those already fall within $[0,1]$) into (4.17) and (4.18) to obtain updated equations of (4.17) and (4.18). Then use Lagrange multiplier to find the values of a_j. If all the a_j s fall within $[0,1]$, then these a_j s are the solutions. Otherwise, repeat the same procedures until all the a_j s are fall within $[0,1]$. (2) Use LINGO software to find values of a_j then use MATLAB to find the increment of the residual sum of squares. We will use the second method for the real data analysis in the next section.

If we apply the constraint of $0 \le a_j \le 1,\ j = 1,2,...,k$ it will affect the increment of the residual sum of squares. Therefore the effect of deleting predictors will be different. In general, the number of deleted predictors under the condition of $0 \le a_j \le 1$ will be less than the number of deleted predictors without this restriction, i.e., $-\infty < a_j < \infty$. This is because in the first step of deleting predictors (see section 4.1), the increment of the residual sum of squares under the condition of $0 \le a_j \le 1$ will be much larger than the increment of the residual sum of squares under the condition of $-\infty < a_j < \infty$. Therefore the F test statistic will be significant under the condition of $0 \le a_j \le 1$ but not significant for the F test statistic under $-\infty < a_j < \infty$. Hence the deletion of predictors will be more conservative under the condition of $0 \le a_j \le 1$.

4.5 The Three Major Factors of Deleting Predictors

Using the deletion method of Mansfield et al. (1977), the order of deleting the predictors and the number of predictors to be deleted will be affected by the following three major factors: (1) determination of principal components (2) after the deletion of a predictor, should we also delete the corresponding column and row in the $X'X$ and $X'Y$ matrices? (3) should we restrict the diagonal elements a_j (of the diagonal matrix A in $\underline{\hat{\beta}_{MGR}}$ = **VAV'** $\underline{\hat{\beta}}$) under the condition of $0 \le a_j \le 1$?

In the principal components regression analysis, we delete those principal components corresponding to smaller eigenvalues, therefore, the number of principal components will be less than the number of predictors. In the next section, we will use real data to discuss whether we should delete those principal components corresponding to smaller eigenvalues because they will affect the results of deleting predictors in the model.

5. Data Analysis

In this section, we will use two examples to compare the results of estimating the MWG parameters and the deletion of predictors under the two different conditions, i.e., $0 \le a_j \le 1$ and $-\infty < a_j < \infty$. The correlation coefficients matrices of these two data sets are given in Appendix D.

5.1 The Pitprop Problem as an Example

5.1.1: The Description of Data

The pitprop data of Jeffers (1967) arise from the study of 13 physical properties of pitprops, X_1 to X_{13}, and their relationship to the compressive strength of the pitprop (Y). A summary of the data of 180 pitprops is given in Jeffers (1967) in the form of pairwise correlations. The 13 variables studied were:

X_1 (TOPDIAM): the top diameter of the pitprop in inches;
X_2 (LENGTH): the length of the pitprop in inches;
X_3 (MOIST): the moisture content of the pitprop, expressed as a percentage of the dry weight;
X_4 (TESTSG): the specific gravity of the timber at the time of the test;
X_5 (OVENSG): the oven-dry specific gravity of the timber;
X_6 (RINGTOP): the number of annual rings at the top of the pitprop;
X_7 (RINGBUT): the number of annual rings at the base of the pitprop;
X_8 (BOWMAX): the maximum bow in inches
X_9 (BOWDIST): the distance of the point of maximum bow from the top of the pitprop in inches;
X_{10} (WHORLS): the number of knot whorls;
X_{11} (CLEAR): the length of clear pitprop from the top of the pitprop in inches;
X_{12} (KNOTS): the average number of knots per whorl;
X_{13} (DIAKNOT): the average diameter of the knots in inches;
Y=Compressive strength

Table 1 in Appendix D gives the coefficients of correlation between each of the 13 variables, one asterisk indicating significance at the 0.05 of probability and two asterisks indicating significance at the 0.01 of probability. The high degree of intercorrelation between the variables is evident from this table.

The two softwares, MATLAB and LINGO, are used to calculate the results given in the tables of this section.

In this example, the ordered eigenvlaues of X'X are: 4.219, 2.378, 1.878, 1.109, 0.910, 0.815, 0.576, 0.440, 0.353, 0.191, 0.0506, 0.0415, and 0.0387. The condition number of the correlation matrix X'X is $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$ =4.219/0.0387=109.02, it indicates there exists a moderate multicollinearity problem.

Table 5.1.1 gives the values of estimated regression coefficients, standard deviations, t test statistic values, p-values, R_j^2 values for full model (all the 13 variables included in the model) and estimated regression coefficients, standard deviations for using backward elimination method.

Table 5.1.1: Summary Statistics for Full Model and Backward Elimination

	Regression on all predictors					Backward Elimination	
Predictor	Regression Coefficient	Standard Deviation	t-value	p-value	R_j^2	Regression Coefficient	Standard Deviation
X_1	-0.4885	1.8802	-0.26	0.3975	0.9239	-0.4921	1.8728
X_2	0.4006	1.9212	0.20	0.4308	0.9272	0.3896	1.8814
X_3	-0.9753	1.7715	-0.55	0.2912	0.9143	-0.9719	1.7690
X_4	0.2925	1.8283	0.16	0.4365	0.9192	0.2854	1.8257
X_5	-0.0822	0.8257	-0.10	0.4602	0.6051	-0.0761	0.8214
X_6	0.1789	1.3659	0.13	0.4483	0.8558	0.2271	1.1991
X_7	0.1351	1.7996	0.075	0.4711	0.9168	--------	--------
X_8	-0.2632	0.7061	-0.37	0.3557	0.4602	-0.2775	0.6746
X_9	-0.0728	0.7524	-0.10	0.4602	0.5245	-0.0745	0.7519
X_{10}	-0.0846	1.1736	-0.07	0.4721	0.8046	--------	--------
X_{11}	0.0967	0.6377	0.15	0.4404	0.3382	0.1190	0.5438
X_{12}	-0.0689	0.6214	-0.11	0.4602	0.3028	-0.0624	0.6128
X_{13}	0.0010	0.6904	0.001	0.5000	0.4353	---------	----------

Note: Backward method eliminated X_7, X_{10}, and X_{13}.

In table 5.1.1, there are five predictors with R_j^2 >0.9 and the other four R_j^2 values are less than 0.5. Therefore, some of the $(VIF)_j$ values are quite large and indicate severe multicollinearity problem. Based on individual test statistic value and p-value, all the variables are not significant. But, the F statistics value for full model shows a significant result. This indicates that some predictors should be deleted. Because of the similarity between MWG variable deletion method and backward elimination method, we list the estimated regression coefficients, standard deviations for using backward elimination in table 5.1.1.

The sum of squares of residual is 0.26914 (for all the 13 predictors), the mean square error with degrees of freedom 166 is 0.0016213. When the three predictors (corresponding to the three smallest eigenvalues) were deleted, the sum of squares of residual increases to 0.31489 and the mean square error increases to 0.0018633

The deletion or not of predictors were decided by comparing the F-statistic value with $F_{0.25}(1,166)$ =1.33. Since there are three major factors to affect the deletion of predictors, we have considered eight possible situations in the data analysis. The eight situations are, MWG method with or without deleting predictors and whether $a_j \in [0,1]$ or not; MWG modified method (see section 4.1) with or without deleting predictors and whether $a_j \in [0,1]$ or not. The estimated regression coefficients and standard error of MWG method can be obtained by using equation (4.12) and (4.14). But the estimated regression coefficients of MWG method with $a_j \in [0,1]$ can't be obtained by using formulas. Therefore, the standard deviations can't be calculated. Hence, we only discuss the procedures and results of deleting predictors.

Case I: Use MWG Method Without Deleting Principal Components

1(a): Without the condition, $a_j \in [0,1]$, i.e., $-\infty < a_j < \infty$

Table 5.1.2: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	0	X_{13}(DIAKNOT)	0.0003	Delete X_{13}
2	0	X_{10}(WHORLS)	0.9032	Delete X_{10}
3	0	X_7(RINGBUT)	0.2660	Delete X_7
4	0	X_9(BOWDIST)	1.4005	Stop (keep X_9)

Table 5.1.3: Estimated regression coefficients of final model

Predictor	Regression Coefficient	Standard Deviation
X_1 (TOPDIAM)	-0.4601	1.8665
X_2 (LENGTH)	0.3366	1.8749
X_3 (MOIST)	-0.9776	1.7463
X_4 (TESTSG)	0.2803	1.8254
X_5 (OVENSG)	-0.0741	0.8169
X_6 (RINGTOP)	0.2646	0.6502
X_7 (RINGBUT)	0	0.0195
X_8 (BOWMAX)	-0.2905	0.6295
X_9 (BOWDIST)	-0.0637	0.7260
X_{10} (WHORLS)	0	0.0083
X_{11} (CLEAR)	0.1117	0.5304
X_{12} (KNOTS)	-0.0699	0.5759
X_{13} (DIAKNOT)	0	0.0029

Note: X_7, X_{10}, and X_{13} are deleted.

Table 5.1.4: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}
0.9697	1.0044	0.9656	0.9686	1.0027	0.9817	0.8898	1.3257	0.3771	20.7292	2.5387	0.9357	-26.27

Case I: Use MWG Method Without Deleting Principal Components

1(b): **With** the condition, $a_j \in [0,1]$

Table 5.1.5: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	0	X_{13} (DIAKNOT)	8.3000	Keep X_{13}

Table 5.1.6: Estimated regression coefficients of final model

Predictor	Regression Coefficient
X_1 (TOPDIAM)	-0.4658
X_2 (LENGTH)	-0.3971
X_3 (MOIST)	-0.9614
X_4 (TESTSG)	0.2866
X_5 (OVENSG)	-0.0836
X_6 (RINGTOP)	0.1478
X_7 (RINGBUT)	0.2256
X_8 (BOWMAX)	-0.1509
X_9 (BOWDIST)	-0.1980
X_{10} (WHORLS)	-0.1239
X_{11} (CLEAR)	0.0922
X_{12} (KNOTS)	-0.0276
X_{13} (DIAKNOT)	0.02569

Note: The estimated regression coefficients of MWG method with $a_j \in [0,1]$ can't be obtained by using formulas. Therefore, the standard deviations can't be calculated.

Table 5.1.7: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}
0.9697	1.0044	0.9656	0.9686	1.0027	0.9817	0.8898	1.3257	0.3771	20.7292	2.5387	0.9357	-26.27

Case II: Use MWG Method and Deleting Three Principal Components Corresponding to the Three Smallest Eigenvalues

II(a): Without the condition, $a_j \in [0,1]$, i.e., $-\infty < a_j < \infty$

Table 5.1.8: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	3	X_{13} (DIAKNOT)	0.1225	Delete X_{13}
2	3	X_9 (BOWDIST)	0.2238	Delete X_9
3	3	X_{12} (KNOTS)	1.8857	Stop(Keep X_{12})

Table 5.1.9: Estimated regression coefficients of final model

Predictor	Regression Coefficient	Standard Deviation
X_1 (TOPDIAM)	-0.0846	0.2846
X_2 (LENGTH)	-0.0369	0.3164
X_3 (MOIST)	-0.4120	0.3216
X_4 (TESTSG)	-0.3178	0.2912
X_5 (OVENSG)	0.1296	0.5639
X_6 (RINGTOP)	0.1909	0.4389
X_7 (RINGBUT)	0.1832	0.2681
X_8 (BOWMAX)	-0.3255	0.6489
X_9 (BOWDIST)	0	0.0024
X_{10} (WHORLS)	-0.0572	0.2454
X_{11} (CLEAR)	0.0971	0.4716
X_{12} (KNOTS)	-0.0573	0.5537
X_{13} (DIAKNOT)	0	0.0012

Note: X_9 and X_{13} are deleted.

Table 5.1.10: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}
0.9879	1.0061	1.0132	0.8720	1.0081	0.9511	0.9398	1.1523	1.1725	0.000	0.0	0.0	0.0

II(b): **With** the condition, $a_j \in [0,1]$

Table 5.1.11: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	3	X_5 (OVENSG)	18.7410	Keep X_5

Table 5.1.12: Estimated regression coefficients of final model

Predictor	Regression Coefficient
X_1 (TOPDIAM)	-0.1098
X_2 (LENGTH)	-0.0580
X_3 (MOIST)	-0.4150
X_4 (TESTSG)	-0.2979
X_5 (OVENSG)	0.2314
X_6 (RINGTOP)	0.1268
X_7 (RINGBUT)	0.1278
X_8 (BOWMAX)	-0.2068
X_9 (BOWDIST)	0.0773
X_{10} (WHORLS)	-0.0702
X_{11} (CLEAR)	0.1004
X_{12} (KNOTS)	0.0518
X_{13} (DIAKNOT)	-0.0327

Note: The estimated regression coefficients of MWG method with $a_j \in [0,1]$ can't be obtained by using formulas. Therefore, the standard deviations can't be calculated.

Table 5.1.13: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}
1.0	1.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0

Case III: Use *Modified* MWG Method Without Deleting Principal Components

III(a): Without the condition, $a_j \in [0,1]$, i.e., $-\infty < a_j < \infty$

Table 5.1.14: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	0	X_{13} (DIAKNOT)	0.0004	Delete X_{13}
2	0	X_{10} (WHORLS)	0.9087	Delete X_{10}
3	0	X_7 (RINGBUT)	0.2692	Delete X_7
4	0	X_9 (BOWDIST)	1.4259	Stop (keep X_9)

Table 5.1.15: Estimated regression coefficients of final model

Predictor	Regression Coefficient	Standard Deviation
X_1 (TOPDIAM)	-0.4862	1.8671
X_2 (LENGTH)	0.3958	1.8750
X_3 (MOIST)	-0.9828	1.7478
X_4 (TESTSG)	0.2866	1.8255
X_5 (OVENSG)	-0.0762	0.8215
X_6 (RINGTOP)	0.2674	0.6515
X_8 (BOWMAX)	-0.2727	0.6641
X_9 (BOWDIST)	-0.0667	0.7262
X_{11} (CLEAR)	0.1142	0.5305
X_{12} (KNOTS)	-0.0708	0.5759

Note: X_7, X_{10}, and X_{13} are deleted.

Table 5.1.16: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}
1.0043	0.9972	0.9942	0.9902	0.9963	0.9657	0.0	0.9813	1.6555	0.0	1.0078	0.00	0.00

Note: Using modified MWG method, there are only 10 predictors left in the study, therefore, matrix A is a 10 by 10 matrix.

III(b): **With** the condition, $a_j \in [0,1]$

Table 5.1.17: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	0	X_{13} (DIAKNOT)	8.3000	Keep X_{13}

Table 5.1.18: Estimated regression coefficients of final model

Predictor	Regression Coefficient
X_1 (TOPDIAM)	-0.4658
X_2 (LENGTH)	-0.3971
X_3 (MOIST)	-0.9614
X_4 (TESTSG)	0.2866
X_5 (OVENSG)	-0.0836
X_6 (RINGTOP)	0.1418
X_7 (RINGBUT)	0.2256
X_8 (BOWMAX)	-0.1509
X_9 (BOWDIST)	-0.1980
X_{10} (WHORLS)	-0.1239
X_{11} (CLEAR)	0.0922
X_{12} (KNOTS)	-0.0277
X_{13} (DIAKNOT)	0.02569

Note: The estimated regression coefficients of MWG method with $a_j \in [0,1]$ can't be obtained by using formulas. Therefore, the standard deviations can't be calculated.

Table 5.1.19: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}
1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0

Case IV: Use *Modified* MWG Method and Deleting Three Principal Components Corresponding to the Three Smallest Eigenvalues

IV(a): Without the condition, $a_j \in [0,1]$, i.e., $-\infty < a_j < \infty$

Table 5.1.20: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	3	X_{13} (DIAKNOT)	0.1225	Delete X_{13}
2	3	X_9 (BOWDIST)	0.1746	Delete X_9
3	3	X_2 (LENGTH)	1.233	Delete X_2
4	2	X_{10} (WHORLS)	0.9556	Delete X_{10}
5	2	X_{12} (KNOTS)	1.1373	Delete X_{12}
6	2	X_{11} (CLEAR)	7.6481	Stop(Keep X_{11})

Table 5.1.21: Estimated regression coefficients of final model

Predictor	Regression Coefficient	Standard Deviation
X_1 (TOPDIAM)	-0.1702	0.6757
X_3(MOIST)	-0.4069	0.3209
X_4(TESTSG)	-0.3144	0.2902
X_5(OVENSG)	0.1309	0.5598
X_6(RINGTOP)	0.2044	0.4193
X_7(RINGBUT)	0.1629	0.2851
X_8(BOWMAX)	-0.3087	0.5999
X_{11}(CLEAR)	0.1119	0.5279

Note: X_2, X_9, X_{10}, X_{12} and X_{13} are deleted.

Table 5.1.22: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_3	a_4	a_5	a_6	a_7	a_8	a_{11}
1.0029	1.0177	0.6621	1.0601	0.8321	0.8909	0.0	0.0

Note: Using modified MWG method, there are only 8 predictors left in the study, therefore, matrix A is a 10 by 10 matrix.

IV(b): **With** the condition, $a_j \in [0,1]$

Table 5.1.23: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	3	X_5(OVENSG)	18.7410	Keep X_5

Table 5.1.24: Estimated regression coefficients of final model

Predictor	Regression Coefficient
X_1 (TOPDIAM)	-0.1098
X_2 (LENGTH)	-0.0580
X_3(MOIST)	-0.4150
X_4(TESTSG)	-0.2979
X_5(OVENSG)	0.2314
X_6(RINGTOP)	0.1268
X_7(RINGBUT)	0.1278
X_8(BOWMAX)	-0.2068
X_9(BOWDIST)	0.0773
X_{10}(WHORLS)	-0.0702
X_{11}(CLEAR)	0.1004
X_{12}(KNOTS)	0.0518
X_{13}(DIAKNOT)	-0.0327

Note: The estimated regression coefficients of MWG method with $a_j \in [0,1]$ can't be obtained by using formulas. Therefore, the standard deviations can't be calculated.

Table 5.1.25: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}
1.0	1.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0

5.1.2 Discussion of Data Analysis

Based on the results presented in the above tables, we obtained the following conclusions:

a. The number of deleted predictors **With** the condition, $a_j \in [0,1]$, is less than the number of deleted predictors **Without** the condition, $a_j \in [0,1]$.

b. **Without** the condition, $a_j \in [0,1]$, and by comparing Table 5.1.2 with Table 5.1.14; Table 5.1.8 with Table 5.1.20, we can show that the modified MWG method is more effective than the original MWG method, i.e., more predictors will be deleted (Mansfield et al. (1977)).

c. **Without** the condition, $a_j \in [0,1]$, Mansfield et al. (1977) thought that deleting principal components corresponding to smaller Eigenvalues will result in deleting more predictors. But Table 5.1.2 and Table 5.1.8 didn't show the result.

d. The predictors deleted by backward method (X_7, X_{10}, and X_{13}) are the same as Case III(a) (see Table 5.1.14). The estimated regression coefficients are also close (see Table 5.1.1 and Table 5.1.15).

5.2 The Acetylene Data as an Example

5.2.1 The Description of Data

This data concerning the percentage of conversion of n-heptane to acetylene and three original explanatory variables (Douglas, C. M., Peck, E. A., and Vining G.G. (2001)). The response variable is Y=percentage of conversion, and the three original explanatory variables are

X_1=T ((contact time-0.0403)/0.03164),
X_2=H ((H_2 (n-Heptane)-12.44)/5.662),
X_3=C ((Temperature-1212.50)/80.623).

The 9 predictors considered in the principal component regression model are:

X_1=T, X_2=H, X_3=C, X_4=TxH (interaction of T and H), X_5=TxC (interaction of T and C), X_6=CxH (interaction of C and H), $X_7 = T^2$, $X_8 = H^2$, $X_9 = C^2$. The Correlation matrix X'X and X'Y are given in Appendix D.

In this example, the ordered eigenvlaues of X'X are: 4.2048, 2.16261, 1.13839, 1.0413, 0.38453, 0.04951, 0.01363, 0.00513, 0.0001. The condition number of the correlation matrix X'X is $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$ =4.2048/0.0001=42048, it indicates there is a severe multicollinearity problem.

The full quadratic model for the acetylene data is

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \varepsilon \quad (5.1)$$

Table 5.2.1 gives the values of estimated regression coefficients, standard deviations, t test statistic values, p-values, R_j^2 values for model (5.1) and estimated regression coefficients, standard deviations for using backward elimination method.

Table 5.2.1: Summary Statistics for model (5.1) and Backward Elimination

	Regression on all predictors					Backward Elimination	
Predictor	Regression Coefficient	Standard Deviation	t-value	p-value	R_j^2	Regression Coefficient	Standard Deviation
X_1	3.2893	2.9785	1.10	0.3117	0.9973	0.3563	0.0360
X_2	-0.5209	0.2037	-2.56	0.0431	0.4041	-0.4884	0.2559
X_3	3.9665	4.0131	0.99	0.3611	0.9985	-------	---------
X_4	-1.8484	0.8588	-2.15	0.0749	0.9678	-1.8660	0.7737
X_5	13.4133	12.475	1.08	0.3236	0.9998	1.5628	1.1945
X_6	-2.0819	0.9199	-2.26	0.0643	0.9719	-2.1686	0.7962
X_7	7.7582	6.4627	1.20	0.2752	0.9994	1.6543	1.0931
X_8	0.3259	0.2747	1.19	0.2803	0.6851	0.2921	0.1893
X_9	4.7475	5.2379	0.91	0.2803	0.9991	-------	----------

Note: X_3 and X_9 are deleted.

In table 5.2.1, there are seven predictors with R_j^2 >0.9 and the other two R_j^2 values are 0.4041 and 0.6851, respectively. Therefore, most of the $(VIF)_j$ values are quite large and indicate severe multicollinearity problem. Based on individual test statistic value and p-value, only X_2, X_4, and X_6 are significant. But, the F statistics value for model (5.1) shows a significant result. This indicates that some predictors should be deleted.

The sum of residual squares is 0.14256 (for all the 9 "principal components"), for degrees of freedom equal to 6, the standard error is 0.02376.

Case I: Use MWG Method Without Deleting Principal Components

1(a): Without the condition, $a_j \in [0,1]$, i.e., $-\infty < a_j < \infty$

Table 5.2.2: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	0	X_3=C	0.0898	Delete X_3
2	0	X_1=T	0.7692	Delete X_1
3	0	$X_8 = H^2$	2.4688	Stop(Keep X_8)

Table 5.2.3: Estimated regression coefficients of final model

Predictor	Regression Coefficient	Standard Deviation
X_1=T	0	0.3743
X_2=H	-0.4833	0.1562
X_3=C	0	0.5192
X_4=TxH	-1.8674	0.6511
X_5=TxC	1.5627	3.9330
X_6=CxH	-2.1702	0.6942
$X_7 = T^2$,	1.6545	0.5514
$X_8 = H^2$	0.2921	0.2114
$X_9 = C^2$	-0.7887	2.1269

Table 5.2.4: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
1.2002	0.9679	0.8692	0.9886	0.9486	0.3777	1.2206	0.3014	0.9648

Case I: Use MWG Method Without Deleting Principal Components

1(b): **With** the condition, $a_j \in [0,1]$

Table 5.2.5: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	0	X_3=C	0.1040	Delete X_3
2	0	$X_8 = H^2$	2.8893	Stop(Keep X_8)

Table 5.2.6: Estimated regression coefficients of final model

Predictor	Regression Coefficient
X_1=T	0.3782
X_2=H	-0.4634
X_3=C	0
X_4=TxH	-1.8654
X_5=TxC	1.5022
X_6=CxH	-2.1950
$X_7 = T^2$,	0.8218
$X_8 = H^2$	0.2405
$X_9 = C^2$	-0.7738

Note: The estimated regression coefficients of MWG method with $a_j \in [0,1]$ can't be obtained by using formulas. Therefore, the standard deviations can't be calculated.

Table 5.2.7: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
0.9999	1.0	1.0	1.0	1.0	1.0	1.0	0.2726	0.9675

Case II: Use MWG Method and Deleting Two Principal Components Corresponding to the Two Smallest Eigenvalues

II(a): Without the condition, $a_j \in [0,1]$, i.e., $-\infty < a_j < \infty$

Table 5.2.8: Estimated regression coefficients of final model

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	2	X_5=TxC	0.1560	Delete X_5
2	2	X_1=T	0.0721	Delete X_1
3	2	$X_9 = C^2$	1.2503	Delete X_9
3	2	$X_8 = H^2$	4.3446	Stop(Keep X_8)

Table 5.2.9: Estimated regression coefficients of final model

Predictor	Regression Coefficient	Standard Deviation
X_1=T	0	0.1265
X_2=H	-0.4916	0.1371
X_3=C	-0.6550	0.0173
X_4=TxH	-1.5763	0.0707
X_5=TxC	0	0.0144
X_6=CxH	-1.8424	0.0616
$X_7 = T^2$,	0.6919	0.0141
$X_8 = H^2$	0.3307	0.1480
$X_9 = C^2$	0	0.3437

Table 5.2.10: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
1.7509	0.8990	0.9004	0.8732	0.9341	2.0711	0.9047	0.0	0.0

Case II: Use MWG Method and Deleting Two Principal Components Corresponding to the Two Smallest Eigenvalues

II(b): **With** the condition, $a_j \in [0,1]$

Table 5.2.11: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	2	X_5=TxC	0.2400	Delete X_5
2	2	X_1=T	0.1003	Delete X_1
3	2	$X_8 = H^2$	4.1997	Stop(Keep X_8)

Table 5.2.12: Estimated regression coefficients of final model

Predictor	Regression Coefficient
X_1=T	0.0
X_2=H	-0.4688
X_3=C	-0.6120
X_4=TxH	-1.4416
X_5=TxC	0.0
X_6=CxH	-1.6969
$X_7 = T^2$,	0.6408
$X_8 = H^2$	0.3186
$X_9 = C^2$	-0.2995

Note: The estimated regression coefficients of MWG method with $a_j \in [0,1]$ can't be obtained by using formulas. Therefore, the standard deviations can't be calculated.

Table 5.2.13: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
0.8398	1.0	1.0	1.0	1.0	1.0	0.7451	0.0	0.0

Case III: Use *Modified* MWG Method Without Deleting Principal Components

III(a): Without the condition, $a_j \in [0,1]$, i.e., $-\infty < a_j < \infty$

Table 5.2.14: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	0	X_3=C	0.0898	Delete X_3
2	0	$X_9 = C^2$	0.0171	Delete X_9
3	0	X_5(OVENSG)	1.9958	Stop(Keep X_5)

Table 5.2.15: Estimated regression coefficients of final model

Predictor	Regression Coefficient	Standard Deviation
X_1=T	0.3565	0.2121
X_2 =H	-0.4883	0.1449
X_4 =TxH	-1.8674	0.6409
X_5 =TxC	1.5627	0.9898
X_6 =CxH	-2.1702	0.6732
$X_7 = T^2$	1.6545	0.9060
$X_8 = H^2$	0.2921	0.1568

Note: Using modified MWG method, there are only 7 predictors left in the study, therefore, matrix A is a 7 by 7 matrix.

Table 5.2.16: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_4	a_5	a_6	a_7	a_8
0.9984	1.0017	0.9998	1.0001	1.0002	0.9758	1.0009

Case III: Use *Modified* MWG Method Without Deleting Principal Components

III(b): **With** the condition, $a_j \in [0,1]$

Table 5.2.17: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	0	X_3=C	0.1040	Delete X_3
2	0	$X_9 = C^2$	0.0176	Delete X_9
3	0	$X_7 = T^2$	2.8421	Stop(Keep X_7)

Table 5.2.18: Estimated regression coefficients of final model

Predictor	Regression Coefficient
X_1=T	0.3753
X_2=H	-0.4641
X_4=TxH	-1.8900
X_5=TxC	1.5329
X_6=CxH	-2.2215
$X_7 = T^2$,	0.8331
$X_8 = H^2$	0.2399

Note: The estimated regression coefficients of MWG method with $a_j \in [0,1]$ can't be obtained by using formulas. Therefore, the standard deviations can't be calculated.

Table 5.2.19: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_4	a_5	a_6	a_7	a_8
0.9984	1.0	1.0	1.0	1.0002	0.9753	1.0

Note: Using modified MWG method, there are only 7 predictors left in the study, therefore, matrix A is a 7 by 7 matrix.

Case IV: Use *Modified* MWG Method and Deleting Two Principal Components Corresponding to the Two Smallest Eigenvalues

IV(a): Without the condition, $a_j \in [0,1]$, i.e., $-\infty < a_j < \infty$

Table 5.2.20: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	2	X_5=TxC	0.1560	Delete X_5
2	2	$X_7 = T^2$	0.0061	Delete X_7
3	2	X_4=TxH	1.6506	Stop(Keep X_4)

Table 5.2.21: Estimated regression coefficients of final model

Predictor	Regression Coefficient	Standard Deviation
X_1=T	0.6191	0.1414
X_2=H	-0.4495	0.1446
X_3=C	-0.3322	0.1375
X_4=TxH	-0.3509	0.5666
X_6=CxH	-0.5719	0.5765
$X_8 = H^2$	0.2936	0.1594
$X_9 = C^2$	-0.7856	0.0831

Table 5.2.22: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_3	a_4	a_6	a_8	a_9
1.0132	0.9987	1.0020	1.0415	1.0005	0.0	0.0

Note: Using modified MWG method, there are only 7 predictors left in the study, therefore, matrix A is a 7 by 7 matrix.

Case IV: Use *Modified* MWG Method and Deleting Two Principal Components Corresponding to the Two Smallest Eigenvalues

IV(b): **With** the condition, $a_j \in [0,1]$

Table 5.2.23: The steps of deleting predictors

Step	# of principals deleted	Predictors to be deleted	F-value	Conclusion
1	2	X_5=TxC	0.0240	Delete X_5
2	2	$X_7 = T^2$	0.0068	Delete X_7
3	2	X_4=TxH	2.5248	Stop(Keep X_4)

Table 5.2.24: Estimated regression coefficients of final model

Predictor	Regression Coefficient
X_1=T	0.4765
X_2=H	-0.4601
X_3=C	-0.2093
X_4=TxH	-0.0549
X_6=CxH	-0.2629
$X_8 = H^2$	0.3190
$X_9 = C^2$	0.3985

Note: The estimated regression coefficients of MWG method with $a_j \in [0,1]$ can't be obtained by using formulas. Therefore, the standard deviations can't be calculated.

Table 5.2.25: The values of diagonal elements a_j of diagonal matrix **A**

a_1	a_2	a_3	a_4	a_6	a_8	a_9
1.0	0.9986	1.0	1.0	0.9075	0.0	0.0

Note: Using modified MWG method, there are only 7 predictors left in the study, therefore, matrix A is a 7 by 7 matrix.

5.2.2 Discussion of Data Analysis

Based on the results presented in the above tables, we obtained the following conclusions:

a. The number of deleted predictors **With** the condition, $a_j \in [0,1]$, is less than the number of deleted predictors **Without** the condition, $a_j \in [0,1]$. This is the same result for the data in the Pitprop data.
b. In the Pitprop example, NO predictors are deleted **With** the condition, $a_j \in [0,1]$. But, in this example there are predictors being deleted (see Table 5.2.5, 5.2.11, 5.2.17, and 5.2.23). One of the reasons is the existing severe multi-collinearity problem.
c. **Without** the condition, $a_j \in [0,1]$, We found the number of predictors deleted by using MWG *modified* method is less than the use of original MWG method. This is a contradiction to the conclusion in Mansfield et al. (1977).
d. In the paper of Mansfield et al. (1977) they mentioned that if we first delete the principal components corresponding to smaller Eigenvalues, it will result in deleting more predictors. We found that is true no matter whether we have the restriction of $a_j \in [0,1]$ or not. (See the comparisons of Tables 5.2.2 with 5.2.; Tables 5.2.5 with 5.2.11; Tables 5.2.14 with 5.2.20, and Tables 5.2.17 with 5.2.23).
e. The predictors deleted by backward method (X_3 and X_9) are the same as Case III(a) (see Table 5.2.14).

6. Conclusions and Suggestions

When there exists a severe multicollinearity problem, Ridge regression is one of the most popular estimation procedures for combating multicollinearity. Biased estimation is used to attain a substantial reduction in variance with an accompanied increase in stability of the regression coefficients. After the data analysis of two examples in section 5, we have the following conclusions:

a. **With** the condition, $a_j \in [0,1]$, the MWG method of deleting predictors became more conservative, i.e., the number of deleted predictors is less than the number deleted without $a_j \in [0,1]$.
b. **Without** the condition, $a_j \in [0,1]$, the values of diagonal elements a_j of diagonal matrix **A** may become very large. Larger values of a_j will increase the variance of MWG estimation and affect the precision of estimation results. For example, In Table 5.1.4, a_{10} =20.729 and a_{13} =-26.173.
c. In the first example, there is moderate multicollinearity problem. If we add the restriction of $a_j \in [0,1]$, the procedures of deleting predictors became conservative. Since we have severe multicollinearity problem in example 2, the effect of deleting predictors almost the same between the restrictions of $a_j \in [0,1]$ and $-\infty < a_j < \infty$.
d. **With** the condition $a_j \in [0,1]$, the severe multicollinearity problem, and the small sample size (n=16), the MSE is larger and all these make the F-statistic not significant. Therefore, more predictors are deleted.

e. There are drawbacks in the two examples we used to do data analyses: (1) the multicollinearity problem is not severe in example 1, therefore the effect of deleting predictors by adding the restriction of $a_j \in [0,1]$ to MWG method is hard to be justified. (2) the sample size is too small in the second example, small sample size increases the MSE and decreases the F-statistics and will affect the number of predictors to be deleted.

f. It is time consuming to use the modified MWG method because we have to reevaluate the model at each step, i.e., recalculate the eigenvalues and eigenvectors; redefine the principal components, etc.

The following are our suggestions:

a. In order to use MWG method or the method proposed in this article to delete predictors, we need a large data set with severe multicollinearity problem so that we one can build up a most appropriate model for data analysis.

b. One should design artificial examples with large sample size and different degree of multicollinearity so that we can study MWG method and our proposed method thoroughly.

APPENDIX A: The Derivation of E($\tilde{\gamma}_s$) and V($\tilde{\gamma}_s$)

Since $\Lambda_s \underline{\hat{\gamma}_s} - \Lambda_s A_s \underline{\hat{\gamma}_s} = V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2 \underline{\hat{\gamma}_s}$ (Mansfield et al. (1977)), we obtained $\Lambda_s A_s \underline{\hat{\gamma}_s} = \Lambda_s \underline{\hat{\gamma}_s} - V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2 \underline{\hat{\gamma}_s}$ it implies

$$\tilde{\gamma}_s = A_s \underline{\hat{\gamma}_s} = [\mathbf{1} - \Lambda_s^{-1} V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2] \underline{\hat{\gamma}_s}.$$

Therefore, $E(\tilde{\gamma}_s) = [\mathbf{1} - \Lambda_s^{-1} V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2] E(\underline{\hat{\gamma}_s}) = [\mathbf{1} - \Lambda_s^{-1} V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2] \underline{\gamma_s}$.

The expected value of $\underline{\tilde{\beta}}$ is $E(\underline{\tilde{\beta}}) = E(V_s \tilde{\gamma}_s) = V_s [\mathbf{1} - \Lambda_s^{-1} V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2] \underline{\gamma_s}$, since $\underline{\hat{\gamma}_s} = (\hat{\gamma}_1, \hat{\gamma}_2, ..., \hat{\gamma}_s)' = \Lambda_s^{-1} W_s' \underline{Y}$, the variance-covariance matrix of $\underline{\hat{\gamma}_s}$ is

$$\mathrm{Var}(\underline{\hat{\gamma}_s}) = \mathrm{Var}(\Lambda_s^{-1} W_s' \underline{Y}) = \Lambda_s^{-1} W_s' W_s \Lambda_s^{-1} \sigma^2$$

But, $W = [W_s : W_{k-s}]$ and $W'W = \Lambda$, it implies

$$W'W = \begin{bmatrix} W_s' \\ W_{k-s}' \end{bmatrix} \begin{bmatrix} W_s & W_{k-s} \end{bmatrix} = \begin{bmatrix} W_s'W_s & W_s'W_{k-s} \\ W_{k-s}'W_s & W_{k-s}'W_{k-s} \end{bmatrix} = \Lambda = \begin{bmatrix} \Lambda_s & 0 \\ 0 & \Lambda_{k-s} \end{bmatrix} \Rightarrow W'W = \Lambda_s.$$

Therefore, $\mathrm{Var}(\underline{\hat{\gamma}_s}) = \mathrm{Var}(\Lambda_s^{-1} W_s' \underline{Y}) = \Lambda_s^{-1} W_s' W_s \Lambda_s^{-1} \sigma^2 = \sigma^2 \Lambda_s^{-1}$.

The variance-covariance of $\tilde{\gamma}_s$ is

$$\begin{aligned}
\mathrm{Var}(\tilde{\gamma}_s) &= \sigma^2 [\mathbf{1} - \Lambda_s^{-1} V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2] \Lambda_s^{-1} [\mathbf{1} - V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2 \Lambda_s^{-1}] \\
&= \sigma^2 [\Lambda_s^{-1} - \Lambda_s^{-1} V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2 \Lambda_s^{-1}] [\mathbf{1} - V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2 \Lambda_s^{-1}] \\
&= \sigma^2 [\Lambda_s^{-1} - \Lambda_s^{-1} V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2 \Lambda_s^{-1} - \Lambda_s^{-1} V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2 \Lambda_s^{-1} + \\
&\qquad \Lambda_s^{-1} V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2 \Lambda_s^{-1} V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2 \Lambda_s^{-1}] \\
&= \sigma^2 [\Lambda_s^{-1} - \Lambda_s^{-1} V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2 \Lambda_s^{-1}]
\end{aligned}$$

Therefore, the variance-covariance of $\underline{\tilde{\beta}} = V_s \underline{\tilde{\gamma}}_s$ is

$$\text{Var}(\underline{\tilde{\beta}}) = \sigma^2 V_s [\Lambda_s^{-1} - \Lambda_s^{-1} V_2' (V_2 \Lambda_s^{-1} V_2')^{-1} V_2 \Lambda_s^{-1}] V_s'$$

APPENDIX B: The Derivation of (4.15)

The residual vector of MWG's estimator is $\tilde{e} = \underline{Y} - \underline{\bar{Y}}\,\mathbf{1} - \mathbf{X}\underline{\tilde{\beta}} = \underline{Y} - \underline{\bar{Y}}\,\mathbf{1} - \mathbf{X} V_s A_s \underline{\hat{\gamma}}_s$ and the prediction vector is $\underline{\bar{Y}}\,\mathbf{1} + \mathbf{X} V_s A_s \underline{\hat{\gamma}}_s$. To show the residual vector is orthogonal to the prediction vector, it implies that $(\underline{Y} - \underline{\bar{Y}}\,\mathbf{1} - \mathbf{X}\underline{\tilde{\beta}})'(\underline{\bar{Y}}\,\mathbf{1} + \mathbf{X} V_s A_s \underline{\hat{\gamma}}_s) = \mathbf{0}$.

$\Leftrightarrow \underline{\hat{\gamma}}_s' A_s V_s' \mathbf{X}' \underline{Y} - \underline{\hat{\gamma}}_s' A_s V_s' \mathbf{X}' \mathbf{X} V_s A_s \underline{\hat{\gamma}}_s = \mathbf{0}$

$\Leftrightarrow \underline{\hat{\gamma}}_s' A_s \Lambda_s \underline{\hat{\gamma}}_s - \underline{\hat{\gamma}}_s A_s \Lambda_s A_s \underline{\hat{\gamma}}_s = \mathbf{0}$

$\Leftrightarrow \underline{\hat{\gamma}}_s' (A_s \Lambda_s - A_s \Lambda_s A_s) \underline{\hat{\gamma}}_s = \mathbf{0}$

APPENDIX C: The proof of (4.16)

The minimization of (4.16) can be rewritten as

$$\text{Min } u_r = \sum_{j=1}^{s} (1-a_j)^2 \lambda_j \hat{\gamma}_j^2$$

$$\text{such that } \sum_{j=1}^{s} a_j (1-a_j) \lambda_j \hat{\gamma}_j^2 = 0$$

$$\sum_{j=1}^{s} a_j V_{tj} \hat{\gamma}_j = 0, \quad t = k-r+1, k-r+2, \ldots, k$$

where $V_2 = \lfloor V_{tj} \rfloor_{r \times s}$ matrix.

Since $\sum_{j=1}^{s} a_j (1-a_j) \lambda_j \hat{\gamma}_j^2 = 0$, $\sum_{j=1}^{s} (1-a_j)^2 \lambda_j \hat{\gamma}_j^2$ can be transformed as $\sum_{j=1}^{s} (1-a_j) \lambda_j \hat{\gamma}_j^2$, the problem is then

$$\text{Max } u_r = \sum_{j=1}^{s} a_j \lambda_j \hat{\gamma}_j^2 = \underline{\hat{\gamma}}_s' \Lambda_s^{1/2} A_s \Lambda_s^{1/2} \underline{\hat{\gamma}}_s$$

$$\text{such that } \sum_{j=1}^{s} a_j (1-a_j) \lambda_j \hat{\gamma}_j^2 = \underline{\hat{\gamma}}_s' \Lambda_s^{1/2} A_s (\mathrm{I} - A_s) \Lambda_s^{1/2} \underline{\hat{\gamma}}_s = 0$$

$$\sum_{j=1}^{s} a_j V_{tj} \hat{\gamma}_j = V_2 A_s \underline{\hat{\gamma}}_s = V_2 \Lambda_s^{1/2} A_s \Lambda_s^{1/2} \underline{\hat{\gamma}}_s = 0, \quad t = k-r+1, k-r+2, \ldots, k$$

Let $c = \Lambda_s^{1/2} \underline{\hat{\gamma}}_s$, $x = A_s c$, $W_2 = V_2 \Lambda_s^{1/2}$, the problem can be rewritten as

$$\text{Max } u_r = c'x$$

$$\text{such that } c'x - x'x = 0 \text{ and } W_2 x = 0$$

Define $L = c'x + \rho (c'x - x'x) + \mu' W_2 x$, where ρ and μ' are Lagrange multipliers. Setting the derivative with respect to x equal to zero gives $x = (c + \rho c + W_2' \mu)/(2\rho)$ (1)

Setting the derivative with respect to ρ equal to zero gives $c'x - x'x = 0$ (2)

Substituting x into (2) we obtained

$c'[(c+\rho c+W_2'\mu)/(2\rho)]-\{[(c+\rho c+W_2'\mu)'(c+\rho c+W_2'\mu)]/(4\rho^2)\}=0$

$\Rightarrow (\rho^2 c'c-c'c-\mu' W_2'c-c'W_2'\mu-\mu' W_2 W_2'\mu)/(4\rho^2)=0$

$\Rightarrow (\rho^2 c'c-c'c-2c'W_2'\mu-\mu' W_2 W_2'\mu)=0$ (3)

Setting the derivative with respect to μ equal to zero gives $W_2 x=0$. Substituting x we obtained $W_2[(c+\rho c+W_2'\mu)/(2\rho)]=0$. Solve for μ,

$\Rightarrow \mu=-(W_2W_2')^{-1}(W_2c+\rho W_2c)$ (4)

Substituting (4) into (3), we obtained

$\rho^2 c'c-c'c-2c'W_2'[-(W_2W_2')^{-1}(W_2c+\rho W_2c)]-[-(W_2W_2')^{-1}(W_2c+\rho W_2c)]' W_2 W_2'$
$[-(W_2W_2')^{-1}(W_2c+\rho W_2c)]=0$

$\Rightarrow \rho^2 c'c-c'c+c'W_2'(W_2W_2')^{-1}W_2c-\rho^2 c'W_2'(W_2W_2')^{-1}W_2c=0$. Substituting the solution of $\rho=1$ into (4), we obtained $\mu=-2(W_2W_2')^{-1}W_2c$.

Substituting the solutions of ρ and μ into (1), we obtained

$$X=c-W_2'(W_2W_2')^{-1}W_2c=[I-W_2'(W_2W_2')^{-1}W_2]c$$

$\Rightarrow A_s c=[I-W_2'(W_2W_2')^{-1}W_2]c$

Since $W_2=V_2\Lambda_s^{1/2}$, $c=\Lambda_s^{1/2}\underline{\hat{\gamma}_s}$, it implies

$A_s\Lambda_s^{1/2}\underline{\hat{\gamma}_s}=\Lambda_s^{1/2}\underline{\hat{\gamma}_s}-\Lambda_s^{-1/2}V_2'(V_2\Lambda_s^{-1}V_2')^{-1}V_2\Lambda_s^{-1/2}\Lambda_s^{1/2}\underline{\hat{\gamma}_s}$

$\Rightarrow A_s\Lambda_s^{1/2}\underline{\hat{\gamma}_s}=\Lambda_s^{1/2}\underline{\hat{\gamma}_s}-\Lambda_s^{-1/2}V_2'(V_2\Lambda_s^{-1}V_2')^{-1}V_2\underline{\hat{\gamma}_s}$

Observe the jth element on both sides of the above equation one can obtain the following equation

$$a_j\lambda_j^{1/2}\hat{\gamma}_j=\lambda_j^{1/2}\hat{\gamma}_j-\lambda_j^{-1/2}v_{2j}'(V_2\Lambda_s^{-1}V_2')^{-1}V_2\underline{\hat{\gamma}_s}$$

$\Rightarrow a_j=1-\{v_{2j}'[V_2\Lambda_s^{-1}V_2']^{-1}V_2\underline{\hat{\gamma}_s}/\lambda_j\hat{\gamma}_j\}$ $\quad j=1,2,...,s$ (4.6)

APPENDIX D: Two Examples

Table 1: The Correlation Matrix of Pitprop Problem

1	0.954	0.364	0.342	-0.129	0.313	0.496	0.424	0.592	0.545	0.084	-0.019	0.134
0.954**	1	0.297	0.284	-0.118	0.291	0.503	0.419	0.648	0.569	0.076	-0.036	0.144
0.364**	0.297**	1	0.882	-0.148	0.153	-0.029	-0.054	0.125	-0.081	0.162	0.22	0.126
0.342**	0.284**	0.882**	1	0.22	0.381	0.174	-0.059	0.137	-0.014	0.097	0.169	0.015
-0.129	-0.118	-0.148*	0.220**	1	0.364	0.296	0.004	-0.039	0.037	-0.091	-0.145	-0.208
0.313**	0.291**	0.153*	0.381**	0.364**	1	0.813	0.09	0.211	0.274	-0.036	0.024	-0.329
0.496**	0.503**	-0.029	0.174*	0.296**	0.813**	1	0.372	0.465	0.679	-0.113	-0.232	-0.424
0.424**	0.419**	-0.054	-0.059	0.004	0.09	0.372**	1	0.482	0.557	0.061	-0.357	-0.202
0.592**	0.648**	0.125	0.137	-0.039	0.211**	0.465**	0.482**	1	0.526	0.085	-0.127	-0.076
0.545**	0.569**	-0.081	-0.014	0.037	0.274**	0.679**	0.557**	0.526**	1	-0.319	-0.368	-0.291
0.084	0.076	0.162*	0.097	-0.091	-0.036	-0.113	0.061	0.085	-0.319**	1	0.029	0.007
-0.119	-0.036	0.220**	0.169*	-0.145*	0.024	-0.232**	-0.357**	-0.127	-0.368**	0.029	1	0.184
0.134	0.144	0.126	0.015	-0.208**	-0.329**	-0.424**	-0.202**	-0.076	-0.291**	0.007	0.184*	1

* indicates significance at the 0.05 of probability.

** indicates significance at the 0.01 of probability.

The above table gives the matrix of X'X of the 13 independent variables (predictors).
X'Y=[-0.419 -0.338 -0.728 -0.543 0.247 0.117 0.11 -0.253 -0.2325 -0.101 -0.055 -0.117 -0.153]'

The 13 predictors are:
X_1=TOPDIAM, X_2=LENGTH, X_3=MOIST, X_4=TESTSG, X_5=OVENSG,
X_6=RINGTOP, X_7=RINGBUT, X_8=BOWMAX, X_9=BOWDIST,
X_{10}=WHORLS, X_{11}=CLEAR, X_{12}=KNOTS, X_{13}=DIAKNOT
Y=Compressive strength

Table 2: The Correlation Matrix of Acetylene Problem

1	0.224	-0.958	-0.132	0.443	0.205	-0.271	0.031	-0.577
0.224	1	-0.24	0.039	0.192	-0.023	-0.148	0.498	0.224
-0.958	-0.24	1	0.194	-0.661	-0.274	0.501	-0.018	0.765
-0.132	0.039	0.194	1	-0.265	-0.975	0.246	0.398	0.274
0.443	0.192	-0.661	-0.265	1	0.323	-0.972	0.126	-0.972
0.205	-0.023	-0.274	-0.975	0.323	1	-0.279	-0.374	0.358
-0.271	-0.148	0.501	0.246	-0.972	-0.279	1	-0.124	0.874
0.031	0.498	-0.018	0.398	0.126	-0.374	-0.124	1	-0.158
-0.577	-0.224	0.765	0.274	-0.972	0.358	0.874	-0.158	1

The above table gives the matrix of X'X of the 9 independent variables (predictors).
X'Y=[0.30218 -0.2296 -0.1977 0.29155 -0.1522 -0.3322 0.22062 0.12149 0.08239]'

The 9 predictors are:
X_1=T, X_2=H, X_3=C, X_4=TxH (interaction of T and H), X_5=TxC (interaction of T and C), X_6=CxH (interaction of C and H), $X_7 = T^2$, $X_8 = H^2$, $X_9 = C^2$.
Y= percentage of conversion

REFERENCES

1. Mason, Gunst and Webster [1975], "Regression Analysis and Problems of Multicollinearity," Communication Statistics., 4(3), p. 277-292.
2. Mansfield, Webster and Gunst [1977], "An Analytic Variable Selection Technique for Principal Component Regression" Appl Stat., 26(1), p.34-40.
3. Graybill, F. A. (1976). Theory and Application of the Linear Model. Boston, Massachusetts:Duxbury Press.
4. Jeffers [1967], "Two Case Studies in the Application of Principal Component Analysis" Appl Stat., 16, p. 225-236.
5. Hoerl and Kennard [1970(a)], "Ridge Regression: Biased Estimation for Nonorthagonal Problems" Technometrics., 12, p. 69-82.
6. Hoerl and Kennard [1970(b)], "Ridge Regression Applications to Nonorthagonal Problems" Technometrics., 12, p. 55-67.
7. G.A.F. Seaber [1975], Linear Regression Analysis, Wiley, New York.
8. Douglas C, Montgomery Elizabeth A, Peck C Ceoffery Vining [2001], Introduction to Linear Regression Analysis, 3rd ed., Wiley, New York.
9. John O. Rawlings, Sastry G. Pantula, David A. Dickey [1998], Applied Regression Analysis – A Research Tool, 2nd ed., Springer, New York.
10. Richard A. Johnson, Dean W. Wichern [1998], Applied Multivariate Statistical Analysis, Prentice Hall.
11. Li-Chun Wang (1990), "Generalized Shrunken Least Squares Estimator", Applied Probability and Statistics (in Chinese), Vol. 6(3), p 219-231.