REGRESSION MODELING USING PRINCIPAL COMPONENTS

Shahar Boneh
Gonzalo R. Mendieta

Follow this and additional works at: http://newprairiepress.org/agstatconference

Part of the Agriculture Commons, and the Applied Statistics Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.

Recommended Citation

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.
REGRESSION MODELING USING PRINCIPAL COMPONENTS

Shahar Boneh and Gonzalo R. Mendieta
Department of Mathematics and Statistics
The Wichita State University

Abstract

In this paper we present a new stepwise method for selecting predictor variables in linear regression models and its application to agricultural data analysis. This method is an extension of principal component regression, and it consists of iteratively selecting original predictor variables one at a time from repeatedly selected subsets of principal components. The reasoning behind the method and its implementation are discussed, and an example of applying the method to agricultural data is given. The example also demonstrates the advantages of the proposed method over some known methods.

Key words: Variable selection, principal components, multicollinearity

1. Introduction

The problem of variable selection in regression models occurs when observations are collected on a large number of variables for exploratory purposes, while the goal is to obtain a model with only a few important predictor variables. Reducing the number of variables is also a common practice when multicollinearity among the predictor variables exists. There are many known variable selection methods, most of which fall in one of two main categories: (1) Exhaustive search methods, based on examining all the possible subsets of predictors and selecting the best subset with respect to some criterion. (2) Systematic selection algorithms such as the forward, backward and stepwise methods.

While these methods often yield reasonably good subsets of predictor variables, they have some disadvantages. The exhaustive search procedures require a large amount of computations, and become very expensive, or even unfeasible in large scale problems. Systematic algorithms, though computationally efficient, sometimes fail to detect the best predictive subset of variables. For a comprehensive survey of variable selection methods, the reader is referred to Miller (1990).

Principal component regression is a well know technique for reducing the estimation variance in regression analysis when multicollinearity is present. Even though it reduces the dimensionality of the space of predictors, this technique has the shortcoming that there is no corresponding reduction in the number of original variables that must be measured. Jeffers (1965, 1967) was the first to argue that principal components can also provide information as to which predictors should be selected. Following Jeffers' work, several methods that use principal components to reduce the number of variables in multiple regression have been suggested by such authors as Jolliffe (1972, 1973), Hawkins (1973), and Mansfield et al. (1977).

The goal of this paper is to present a new method of using principal component regression for the selection of predictor variables. This proposed method is stepwise in nature, and it is based on repeated inversions from principal components to the
original variables. The method appears to be particularly efficient for data sets with a high degree of multicollinearity. In section 2 we give a brief review of principal component regression. In section 3 the new selection method is described. In section 4 we discuss the reasoning and implementation of the method, and show that it is as computationally efficient as the widely used stepwise selection method. In section 5 we illustrate an application of the method to a data set concerning pitprop strength, and compare it to other methods.

2. Principal Component Regression

We consider the standard linear regression model \( Y = X\beta + \varepsilon \), where \( Y \) is an \( n \times 1 \) vector of responses, \( X = [X_1, ..., X_p] \) is an \( n \times p \) full rank matrix of predictor variables, \( \beta \) is a \( p \times 1 \) vector of unknown parameters, and \( \varepsilon \) is an \( n \times 1 \) vector of uncorrelated and normally distributed random errors with expectation 0 and common variance \( \sigma^2 \), i.e., \( \varepsilon \sim N_n(0, \sigma^2 I_n) \). Thus, \( x_{ij} \) is the \( i \)-th observation on the \( j \)-th predictor variable. Throughout this paper it is assumed without loss of generality that all the variables are standardized, i.e.,

\[
\sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1 \quad \text{for all } j, \quad \text{and} \quad \sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} y_i^2 = 1.
\]

Thus, \( X^T X \) is the correlation matrix of \( X \), and \( X^T Y \) is the correlation vector between \( X \) and \( Y \).

Let \( \hat{\beta} \) denote the least squares estimator of \( \beta \). Then \( \hat{\beta} = (X^T X)^{-1} X^T Y \), and it has the properties: (1) \( \hat{\beta} \) is unbiased, i.e., \( E(\hat{\beta}) = \beta \). (2) \( \text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \).

Let \( V = [V_1, ..., V_p] \) be the matrix of size \( p \times p \) whose columns are the normalized eigenvectors of \( X^T X \), and let \( \lambda_1, ..., \lambda_p \) be the corresponding eigenvalues.

Let \( W = [W_1, ..., W_p] = XV \). Then \( W_j = XV_j \) is the \( j \)-th sample principal component of \( X \). Some important properties of principal components are:

1. \( V^T V = VV^T = I_p \), i.e., the matrix \( V \) is orthonormal.
2. \( W^T W = \Lambda \), where \( \Lambda = \text{diag}(\lambda_1, ..., \lambda_p) \), i.e., \( W \) is orthogonal and \( |W_j| = \sqrt{\lambda_j} \).
3. \( X = WV^T \) and \( X_j = \sum_{k=1}^{p} v_{jk} W_k \).

Now the regression model can be written as:

\[
Y = X\beta + \varepsilon = XVV^T \beta + \varepsilon = W\gamma + \varepsilon, \quad \text{where} \quad \gamma = V^T \beta.
\]

Under this formulation, the least squares estimator of \( \gamma \) is

\[
\hat{\gamma} = (W^T W)^{-1} W^T Y = \Lambda^{-1} W^T Y,
\]

and hence, the principal component estimator of \( \beta \) is defined by

\[
\hat{\beta} = V\hat{\gamma} = V\Lambda^{-1} W^T Y.
\]
the estimation of $\gamma$. Thus,
\[
\hat{\gamma} = (W_s^T W_s)^{-1} W_s^T Y = A_s^{-1} W_s^T Y,
\]
and hence, (2.2) can be rewritten as
\[
\tilde{\beta} = V_s A_s^{-1} W_s^T Y. \tag{2.3}
\]
It can be shown that \(\text{var}(\tilde{\beta}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} \) and \(\text{var}(\hat{\beta}) = \sigma^2 \sum_{j=1}^{s} \frac{1}{\lambda_j} \), where \(\text{var}(\tilde{\beta})\) and \(\text{var}(\hat{\beta})\) refer to the traces of the corresponding variance-covariance matrices. It follows that \(\text{var}(\tilde{\beta}) < \text{var}(\hat{\beta})\), but on the other hand, \(E(\tilde{\beta}) \neq \beta\), that is, the principal component estimator is biased. For this reason, principal component regression belongs to the class of the so-called 'biased regression methods'. It can also be shown that the mean squared error of \(\hat{\beta}\), (the variance plus the squared bias) is given by
\[
\text{MSE}(\hat{\beta}) = \sigma^2 \sum_{j=1}^{s} \frac{1}{\lambda_j} + \sum_{j=s+1}^{p} \gamma_j^2. \tag{2.4}
\]

The main question in principal component regression is which components to delete. From the expression of \(\text{var}(\tilde{\beta})\) it can be seen that the maximal reduction of variance is achieved by deleting the principal components associated with the smallest eigenvalues, but it is also desirable to keep principal components with large coefficients to avoid a large bias. Unfortunately, these two criterions do not always coincide. Kung and Shariff (1980) as well as Jolliffe (1982) give examples in which components with small eigenvalues are important in the regression model. Several approaches can be taken in determining which principal components should be deleted. The most common ones are to delete components so as to minimize \(\text{MSE}(\hat{\beta})\), as given in equation (2.4), or to use the method of latent root regression. Elaborate discussions of principal component regression can be found in Coxe (1982, 1984), Jolliffe (1986), and Jackson (1991).

3. The Selection Method

In this section we describe the selection method step by step. The reasoning and some computational details will be discussed in section 4. Prior to starting the selection, a significance level $\alpha$ has to be set. The level of 0.05 is usually adequate.

1. Selection of the first variable
1.1. Obtain the principal components, \([W_1, ..., W_p]\), of \([X_1, ..., X_p]\).
1.2. Regress the response $Y$ on \([W_1, ..., W_p]\), to fit the model $Y = W\gamma + \epsilon$.
1.3. Let $W_{(s)}$ be the subset of \([W_1, ..., W_p]\) containing the principal components for which the regression coefficient $\hat{\gamma}_j$ is significant at level $\alpha$. The significance of each $\hat{\gamma}_j$ is determined by the usual t-test. (In practice, $W_{(s)}$ should be modified to accommodate for principal components with small eigenvalues. More details are given in section 4).
1.4. If $W_{(s)}$ is an empty set, then the selection process is terminated with the conclusion that no predictor variables should be included in the model. Otherwise, let $\text{SSE}_j$, $j = 1, ..., p$, denote the error sum of squares when $X_j$ is regressed on $W_{(s)}$. 

New Prairie Press
The first predictor variable selected is the one for which $\text{SSE}_j$ is minimal. Without loss of generality let us assume that $X_1$ is selected.

2. Selection of the second variable.

2.1. Regress $(X_2, ..., X_p)$ on $X_1$, and let $e_j$ ($j = 2, ..., p$) denote the vector of standardized residuals obtained from regressing $X_j$ on $X_1$.

2.2. Obtain the principal components $[W'_2, ..., W'_p]$, of $[e_2, ..., e_p]$.

2.3. Regress the response $Y$ on $[W'_2, ..., W'_p]$, to fit the model $Y_i = \sum_{j=2}^{p} \delta_j W'_j + \epsilon_i$.

2.4. Let $W'(s)$ denote the subset of $[W'_2, ..., W'_p]$ containing those principal components for which the regression coefficient $\hat{\delta}_j$ is significant at level $\alpha$. (As in step 1.3, this criterion should be modified if the corresponding eigenvalue is small).

2.5. If $W'(s)$ is empty, then the selection process terminates with the conclusion that only the variable $X_1$ should be included in the model. Otherwise, let $\text{SSE}_j$, $j = 2, ..., p$, denote the error sum of squares when $e_j$ is regressed on $W'(s)$. The second predictor variable selected is the one for which $\text{SSE}_j$ is minimal. Without loss of generality again, assume that $X_2$ is selected.

3. Verification of $X_1$.

After selecting the second variable, it is checked whether or not the first selected variable should stay in the model. This is done as follows:

3.1. Regress $X_1$ on $X_2$, and obtain the standardized residuals vector, $e^*$.

3.2. Regress $Y$ on $e^*$, to fit the simple regression model $Y_i = \theta + \eta e^*_i + \epsilon_i$. (In general, $Y$ is to be regressed on the principal components of the standardized residuals. Since in this step there is only one vector of residuals, the principal component will be the vector itself).

3.3. If $\hat{\eta}$ is significant at level $\alpha$, we conclude that $X_1$ should stay in the model. Otherwise, $X_1$ is dropped.

We continue to select additional variables in a similar manner, according to the following general steps:

I. Let $X_{(s)}$ and $X_{(r)}$ be respectively the sets of the previously selected and remaining predictor variables.

II. Regress each variable in $X_{(r)}$ on all the variables in $X_{(s)}$ and obtain the corresponding vectors of standardized residuals $\{e_j, j \in (r)\}$.

III. Obtain the principal components $[W]$ of the $\{e_j\}$ vectors, and regress $Y$ on $[W]$.

IV. Let $W'(s)$ be the subset of $[W]$ containing the principal components for which the regression coefficients are significant at level $\alpha$. (With $\epsilon$ possible modification for small eigenvalues).

After the selection of each variable, the previously selected variables are verified according to the following general steps, which are essentially the reverse of the selection steps:

I. Let $X_k$ denote the most recently selected variable, i.e., the one which was selected...
in the current step, and let $X_{(c)}$ denote the set of the previously selected variables. Regress each of the variables in $X_{(c)}$ on $X_k$ and obtain the corresponding standardized residuals $e_{(c)}$.

II. Obtain the principal components $W_{(c)}$ of $e_{(c)}$ and regress $Y$ on $W_{(c)}$.

II. If all of the regression coefficients are significant at level $\alpha$ (after the modification for small eigenvalues), we conclude that all the variables in $X_{(c)}$ should stay in the model.

III. Otherwise, we conclude that at least one variable in $X_{(c)}$ should be dropped from the model. To determine which variable(s) we drop, let $W_{(n)}$ be the subset of $W_{(c)}$ containing the principal components with the non-significant coefficients. Regress each variable in $X_{(c)}$ on $W_{(n)}$, and obtain $\text{SSE}_j$, $j \in (c)$.

IV. The variable in $X_{(c)}$ with the minimal $\text{SSE}_j$ is then dropped from the model. If the minimum $\text{SSE}_j$ is shared by several variables, then all of them are to be dropped from the model.

After the verification of the previously selected variables is completed, we proceed to select the next variable. The process terminates as soon as there are no significant principal components in the selection step.

4. Discussion

The proposed method is an extension of principal component regression, hence it is based on the underlying idea of utilizing the orthogonality of principal components as opposed to the non-orthogonality of the original variables. Thus, the main two features of our method are: (1) Application of an inferential criterion to select principal components, rather than directly selecting original variables. (2) Inversion from the subspace of the selected principal components to a subspace of the original variables.

The inversion from one subspace to the other is based on the easily verified fact that $\text{SSE}_j$ (as defined in the general selection step IV) is a norm on $\mathbb{R}^n$. Hence, it can be used as a measure of distance between the remaining variable $X_j$ and the linear space spanned by the selected principal components. Therefore, we select the predictor variable with the smallest $\text{SSE}$. The validation of previously selected variables is useful for the same reason as in the usual stepwise regression.

We now comment on the problem of small eigenvalues when determining $W_{(s)}$. Mason and Gunst (1985) point out that the t-statistic used for testing the significance of a given principal component is proportional to its eigenvalue $\lambda$, hence the power of the test is reduced when $\lambda$ is small. In such a case it is advised to modify the test criterion. A computational study of Mason and Gunst (1985) shows that the main loss of power occurs when $\lambda$ falls below 0.1. Therefore, we recommend the following modification: Whenever $\lambda$ is less than 0.1, the selection criterion be altered to select principal components for which

$$\frac{\hat{\gamma}}{\lambda^2 \text{MSE}} > 1.$$ (4.1)

Note that for any principal component, the left hand side of (4.1) equals to the ratio of increase in squared bias result of deletion of that component, resulting from deletion of that component. Therefore, the above recommendation implies that principal components with eigenvalues less than 0.1 should be selected by the ratio (4.1), rather than by the usual t-test. This modification turned out to work well, but by no means we claim that this is the
only good modification.

Regressing each time the remaining variables on the previously selected variables partitions the variable space into two orthogonal subspace. Consequently, the significance test to determine the selection in the next step is adjusted for the effect of the previously selected variables. Also, as a result of the orthogonal partition, the number of principal components selected at each step provides information about the effective dimension of the remaining variable space. Hence the stopping rule employed by this method correctly points out the number of predictor variables that should be included in the model. In light of this, one would expect that the number of principal components selected should be non-increasing as variables enter the model. This indeed was the case whenever the method was applied (see for instance, the example in Section 5), hence we stipulate the following conjecture:

Conjecture: The number of significant principal components is non-increasing from the selection of one variable to the next.

The main strength of the method is in handling data sets which involve a complex structure of multicollinearity. In such cases the stepwise selection method tends to over-select as well as enter correlated variables. In other words, when the correlation structure among the predictor variables is highly convoluted, the stepwise method may fail to sufficiently eliminate the multicollinearity in the data. Numerous trials on ill-conditioned data showed that our method produced models with a better predictive capability than the ones produced by the stepwise method, yet, a reasonably good fit was preserved. This type of situation is illustrated in the example in Section 5.

The procedure can be implemented entirely by manipulations of the correlation matrix only, without reference to the data itself. This feature enables the method to compete very well in terms of computational efficiency. To illustrate how the procedure may be carried out from the correlation matrix, we now give some of the formulas that can be applied in the various steps. The proofs are based on well known properties of regression and principal components, and are omitted for the sake of brevity. These formulas also show that the type of manipulations required by our method are computationally similar to the ones required by the stepwise selection method.

(1) To select principal components, (steps 1.3, 2.4 and in general, step III), we test the hypothesis \( H_0: \gamma_j = 0 \). The standard t-test is used, and it is implemented as follows:

\[
\text{Reject } H_0 \text{ if } \frac{V_j^T (X^T Y)}{\sqrt{1 - (X^T Y)\Lambda \Lambda^T (X^T Y)}} > t_{n-p, \alpha/2}. \tag{4.2}
\]

In the general step, the matrix \( X \) is replaced by the appropriate residual matrix.

(2) To select the next predictor variable based on the selected principal components, \( \text{SSE}_j, \ j \in (r), \) (general step IV) is calculated by:

\[
\text{SSE}(X_j) = \sum_k V_{jk}^2 \lambda_k. \tag{4.3}
\]
(3) Let \( E_j \) be the vector of residuals when regressing \( X_j \) \((j \in X(r))\) on \( X(s) \) in any step, and let \( e_j = \frac{E_j}{S(E_j)} \) = The j-th standardized residual.

Denote \( e = \{ e_j, j \in (r) \} \).

It is clear that all we need are \( e^T e \) and \( e^T Y \). These are given by:

\[
e^T e = \left( \frac{A_{ij}}{B_{ij}} \right), \quad i, j = 2, \ldots, p, \quad \text{and} \quad e^T Y = \left( \frac{C_j}{\sqrt{A_{jj}}} \right), \quad j = 2, \ldots, p, \tag{4.4}
\]

where \( A_{ij} = (X_i^T X_j) - (X_i^T X(s)(X(s)^T X(s))^{-1}(X(s)^T X_j)), \)

\( B_{ij} = \sqrt{A_{ii} \cdot A_{jj}} \)

and \( C_j = (Y^T X_j) - (Y^T X(s)(X(s)^T X(s))^{-1}(X(s)^T X_j)). \)

5. An Example

In this section we illustrate an application of the variable selection method to agricultural data. The data set we consider is known as the Pitprop Data, and it was taken from Jeffers (1967). The data was collected by the Forestry commission in cooperation with the Forest Products Research Laboratory. It is concerned with a study of the compressive strength of pitprops cut from home-grown timber, with the object of determining whether or not such pitprops are sufficiently strong for use in mines. Measurements were obtained from a carefully designed sampling scheme, props being taken at random from a defined population of species, size class, geographical region, and type of seasoning. The method of testing compressive strength consisted of positioning the props vertically between two spherically seated platens, and applying load at the rate of 200 lb. per sq. in. per minute, until failure occurred. The maximum load was then recorded. In addition to the maximum compressive strength, thirteen other variables which represent various physical properties, were measured on each prop that was tested. Table 1 gives a list of these variables.

The data we are using in this example was obtained from a sample of 180 props of Corsican pine from East Anglia, including both seasoned and unseasoned props. Table 2 gives the coefficients of correlation between each of the 13 variables and between each of the variables and the response (compressive strength). An asterisk indicates significance at level 0.05. The high degree of intercorrelation between the predictor variables is evident from this table. In particular, it is interesting to note that many variables are correlated with the length of the prop, and with the number of annual rings at the base of the prop. It is clear that using all the 13 variables in a regression model is highly inadequate.

The variable selection method described in Section 3 was applied to this data set with \( \alpha = 0.05 \). Table 3 gives a summary of the selection process. The variables selected are: \( X_2, X_3, X_5, X_6, \) and \( X_8 \), namely, the length of the prop, the moisture content of the prop, the oven-dry specific gravity of the timber, the number of annual rings at the top of the prop, and the maximum bow.

Notice that the coefficients of correlation between the selected variables are typically small, the highest one being 0.419 (between \( X_2 \) and \( X_8 \)). The multiple \( R^2 \) for the selected model is 0.688. It is also worth mentioning that these selected
variables agree quite well with knowledge of the physical properties of such props.

Upon selecting the subset of variables, some additional data were artificially generated from the multivariate normal distribution with the correlation matrix in Table 2, thus playing the role of "future observations". The selected model was then fitted to these data. The fits which were obtained were very satisfactory with R^2 values exceeding 0.90.

For comparison purposes, we applied the all-subset regressions procedure for this data. The results are summarized in Table 4. For each subset size, Table 4 gives the best subset with respect to R^2 and C_p. It can be seen from the table, that a subset of five variables is quite appropriate, and that the R^2 value achieved by our selected model is close to the best possible R^2 with five variables. Thus, a reasonably good fit was obtained by the new method.

We also applied the stepwise selection method to this data (at the same alpha levels). The stepwise method selected the variables \{X_1, X_2, X_3, X_6, X_8, X_{11}\}. As seen from Table 4, this subset of variables has the highest R^2 among the subsets of six variables, but it is certainly excessive, since the stepwise method selected both X_1 and X_2 which are highly correlated with each other.

To conclude the example, let us compare with the selection method suggested by Mansfield et al. (1977), which is also based on principal components. Applying Mansfield's method to the pitprop data gives the following variables: \{X_1, X_3, X_4, X_5, X_6, X_7, X_8, X_{11}\}. We first note that the variables X_5, X_6, X_7, X_8, X_11 are common to the two methods. Furthermore, since Corr(X_1, X_2) = 0.95, X_1 and X_2 are practically interchangeable. Thus, Mansfield's method selected roughly the same variables as our method did, but with the additional three variables, X_3, X_4, and X_7. The correlation matrix (Table 2) shows that Corr(X_3, X_4) = 0.88 and Corr(X_6, X_7) = 0.81, that is, X_4 and X_7 are highly correlated with other selected variables, hence including them in the model is inadequate. We also note that Corr(X_{11}, Y) = 0.05, hence, X_{11} cannot be regarded as an important variable, and there is no apparent reason to include it in the model. Thus, besides selecting too many variables, Mansfield's method made poor choices as well. We would like to emphasize that it is not our goal to put down other selection methods, but rather to demonstrate that our proposed method is sensible and works well even for 'problematic' data, where some other methods may yield misleading results.

6. Concluding Remarks

The effectiveness of the new variable selection method presented in this paper stems from combining the advantages of principal component regression and the general stepwise approach. The development of this method originated from practical need. The authors encountered several data sets in the areas of agriculture and industrial engineering in which the traditional variable selection methods did not perform in a satisfactory manner, mainly due to complex multicollinearities. On the other hand, it seemed that even though the usefulness of principal component regression has been recognized for a long time, its potential applicability to model selection problems has not been fully realized. It was demonstrated in this paper that the new method can provide a useful supplement to the existing selection methods.

Using formulas (4.2) - (4.4), the selection procedure can be programmed without much difficulty in any programming language that can operate on matrices. The authors have already developed a program in the statistical language S-PLUS that carries out the entire selection procedure. The total amount of computations required by our method is equivalent to the amount of computations required by the stepwise selection and other similar methods.
For the sake of presentation we focused on the linear regression model. However, it should noted that this method can be extended to apply to more general models as well. Extensions and further applications will appear future publications. It is our hope that the procedure will become widely accepted and incorporated in statistical software packages.

References:


Table 1 - The Variables

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>X_1</td>
<td>TOPDIAM</td>
<td>the top diameter of the prop;</td>
</tr>
<tr>
<td>X_2</td>
<td>LENGTH</td>
<td>the length of the prop;</td>
</tr>
<tr>
<td>X_3</td>
<td>MOIST</td>
<td>the moisture content of the prop, as a percentage of the dry weight;</td>
</tr>
<tr>
<td>X_4</td>
<td>TESTSG</td>
<td>the specific gravity of the timber at the time of the test;</td>
</tr>
<tr>
<td>X_5</td>
<td>OVENSG</td>
<td>the oven-dry specific gravity of the timber;</td>
</tr>
<tr>
<td>X_6</td>
<td>RINGTOP</td>
<td>the number of annual rings at the top of the prop;</td>
</tr>
<tr>
<td>X_7</td>
<td>RINGBUT</td>
<td>the number of annual rings at the base of the prop;</td>
</tr>
<tr>
<td>X_8</td>
<td>BOWMAX</td>
<td>the maximum bow;</td>
</tr>
<tr>
<td>X_9</td>
<td>BOWDIST</td>
<td>the distance of the point of maximum bow from the top of the prop;</td>
</tr>
<tr>
<td>X_10</td>
<td>WHORLS</td>
<td>the number of knot whorls;</td>
</tr>
<tr>
<td>X_11</td>
<td>CLEAR</td>
<td>the length of clear prop from the top of the prop;</td>
</tr>
<tr>
<td>X_12</td>
<td>KNOTS</td>
<td>the average number of knots per whorl;</td>
</tr>
<tr>
<td>X_13</td>
<td>DIAKNOT</td>
<td>the average diameter of the knots;</td>
</tr>
</tbody>
</table>

All the measurements of length are in inches.

Table 2 - Coefficients of correlation between physical properties of props

<table>
<thead>
<tr>
<th></th>
<th>STRENGTH</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>RINGBUT</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>STRENGTH</td>
<td>TOPO1AM -0.419*</td>
<td>TOPO1AM</td>
<td></td>
<td></td>
<td></td>
<td>BOWMAX 0.372*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LENGTH</td>
<td>-0.338*</td>
<td>0.954*</td>
<td>LENGTH</td>
<td></td>
<td></td>
<td>BOWDIST 0.465*</td>
<td>0.482*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOIST</td>
<td>-0.728*</td>
<td>0.364*</td>
<td>0.297*</td>
<td>MOIST</td>
<td></td>
<td>WHORLS 0.679*</td>
<td>0.557*</td>
<td>0.526*</td>
<td>WHORLS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TESTSG</td>
<td>-0.543*</td>
<td>0.342*</td>
<td>0.284*</td>
<td>0.882*</td>
<td>TESTSG</td>
<td>CLEAR -0.113</td>
<td>0.061</td>
<td>0.085</td>
<td>-0.319*</td>
<td>CLEAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OVENSG</td>
<td>0.247*</td>
<td>-0.129</td>
<td>-0.118</td>
<td>-0.148</td>
<td>0.220*</td>
<td>OVENSG 0.247*</td>
<td>-0.129</td>
<td>-0.118</td>
<td>-0.148</td>
<td>0.220*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RINGTOP</td>
<td>0.117</td>
<td>0.313*</td>
<td>0.291*</td>
<td>0.153</td>
<td>0.381*</td>
<td>RINGTOP 0.117</td>
<td>0.313*</td>
<td>0.291*</td>
<td>0.153</td>
<td>0.381*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RINGBUT</td>
<td>0.110</td>
<td>0.496*</td>
<td>0.503*</td>
<td>-0.029</td>
<td>0.174</td>
<td>RINGBUT 0.110</td>
<td>0.496*</td>
<td>0.503*</td>
<td>-0.029</td>
<td>0.174</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOWMAX</td>
<td>-0.253*</td>
<td>0.424*</td>
<td>0.419*</td>
<td>-0.054</td>
<td>-0.059</td>
<td>BOWMAX -0.253*</td>
<td>0.424*</td>
<td>0.419*</td>
<td>-0.054</td>
<td>-0.059</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOWDIST</td>
<td>-0.235*</td>
<td>0.592*</td>
<td>0.648*</td>
<td>0.125</td>
<td>0.137</td>
<td>BOWDIST -0.235*</td>
<td>0.592*</td>
<td>0.648*</td>
<td>0.125</td>
<td>0.137</td>
<td>-0.039</td>
<td>0.211*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WHORLS</td>
<td>-0.101</td>
<td>0.545*</td>
<td>0.569*</td>
<td>-0.081</td>
<td>-0.014</td>
<td>WHORLS -0.101</td>
<td>0.545*</td>
<td>0.569*</td>
<td>-0.081</td>
<td>-0.014</td>
<td>0.037</td>
<td>0.274*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLEAR</td>
<td>-0.055</td>
<td>0.084</td>
<td>0.076</td>
<td>0.162</td>
<td>0.097</td>
<td>CLEAR -0.055</td>
<td>0.084</td>
<td>0.076</td>
<td>0.162</td>
<td>0.097</td>
<td>-0.091</td>
<td>-0.036</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KNOTS</td>
<td>-0.117</td>
<td>-0.019</td>
<td>-0.036</td>
<td>0.220*</td>
<td>0.169</td>
<td>KNOTS -0.117</td>
<td>-0.019</td>
<td>-0.036</td>
<td>0.220*</td>
<td>0.169</td>
<td>-0.145</td>
<td>0.024</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIAKNOT</td>
<td>-0.153*</td>
<td>0.134</td>
<td>0.144</td>
<td>0.126</td>
<td>0.015</td>
<td>DIAKNOT -0.153*</td>
<td>0.134</td>
<td>0.144</td>
<td>0.126</td>
<td>0.015</td>
<td>-0.208*</td>
<td>-0.329*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

An asterisk indicates significance at level 0.05.
Table 3 - Summary of Variable Selection

<table>
<thead>
<tr>
<th>Step No.</th>
<th>Number of principal components selected</th>
<th>Variable selected</th>
<th>Variable dropped</th>
<th>Current subset of variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>X₄</td>
<td>--</td>
<td>{ X₄ }</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>X₅</td>
<td>--</td>
<td>{ X₄, X₅ }</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>X₃</td>
<td>X₄, X₅*</td>
<td>{ X₃ }</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>X₅</td>
<td>--</td>
<td>{ X₃, X₅ }</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>X₆</td>
<td>--</td>
<td>{ X₃, X₅, X₆ }</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>X₈</td>
<td>--</td>
<td>{ X₃, X₅, X₆, X₈ }</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>X₂</td>
<td>--</td>
<td>{ X₂, X₃, X₅, X₆, X₈ }</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>Process terminates</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* In the verification part of step 3, one of the two principal component was not significant. The two residual vectors regressed on that component produced the same SSE. Thus, according to verification step IV, both were dropped.

Table 4 - Summary of All-Subset Regressions

<table>
<thead>
<tr>
<th>Subset size</th>
<th>Variables</th>
<th>R-square</th>
<th>C_p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X₃</td>
<td>0.530</td>
<td>113.89</td>
</tr>
<tr>
<td>2</td>
<td>X₃, X₈</td>
<td>0.616</td>
<td>63.04</td>
</tr>
<tr>
<td>3</td>
<td>X₃, X₆, X₈</td>
<td>0.684</td>
<td>22.84</td>
</tr>
<tr>
<td>4</td>
<td>X₃, X₆, X₈, X₁₁</td>
<td>0.695</td>
<td>18.27</td>
</tr>
<tr>
<td>5</td>
<td>X₁, X₃, X₆, X₈, X₁₁</td>
<td>0.705</td>
<td>13.66</td>
</tr>
<tr>
<td>6</td>
<td>X₁, X₂, X₃, X₆, X₈, X₁₁</td>
<td>0.716</td>
<td>9.14</td>
</tr>
<tr>
<td>7</td>
<td>X₁, X₂, X₃, X₄, X₆, X₈, X₁₁</td>
<td>0.721</td>
<td>8.03</td>
</tr>
<tr>
<td>8</td>
<td>X₁, X₂, X₃, X₄, X₆, X₈, X₁₁, X₁₂</td>
<td>0.724</td>
<td>7.92</td>
</tr>
</tbody>
</table>