ABSTRACT

Stepwise methods are well known to be suboptimal at determining "best" subset regressions. Computing all possible regressions to identify "best" subsets becomes a formidable task for more than fifteen variables. However, recent advances in subset selection methods now make their application for isolating "best" regressions feasible for studies containing as many as 50 or more variables. This paper describes a new SAS procedure, named BESTREG, which implements the most recent version of the Furnival-Wilson "leaps and bounds" algorithm for "best" subset selection in regression analysis. This procedure, intended primarily as an exploratory tool for mathematical model development, provides summary statistics for the "best" and alternative "next best" regressions of all sizes. In addition, it provides graphic facilities to assist the user in identifying the overall "best" model and to assess the strength of competing models.

1. INTRODUCTION

The most commonly used tool for mathematical model building is regression analysis. In many cases the model considered is linear and takes the form

\[ Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p \]

where \( Y \) is the dependent variable, the \( X \)'s are a set of \( p \) independent variables, and the \( b \)'s are estimated parameters. Model building is usually an exploratory process; many \( X \)'s are initially offered as candidates to be evaluated for inclusion in a final prediction model.

In the process of developing the model two opposing principles are usually engaged. For purposes of prediction the model generally should include as few \( X \)'s as possible. On the other hand, in many cases they were the "next best". This process thus becomes that of determining the smallest combination of \( X \)'s which provide a good prediction model. This process is referred to as "isolating the best regression equation".

The problem of isolating the best regression is two-fold. Not only must the optimal subset size be determined, but also for a given subset size, the best combination of variables must be identified. A sure-fire method for finding the "best" subset is to compute and enumerate all regression subsets. However, this is only feasible for a small number of variables. From among \( k \) candidates there exist \( C_k \) subsets of size \( p \) or \( 2^k - 1 \) total possible subsets. For \( k = 60 \), approximately \( 1.15 \times 10^{15} \) regression subsets exist.

Most criteria used for determining the best combination of variables of a given subset size optimize some function of the residual sum-of-squares (RSS) \([1]\). The function is optimized when RSS is minimized or equivalently when \( R^2 \), the coefficient of determination, is maximized. Even with today's most advanced computers, computing all possible regressions becomes a prohibitive task for moderate to large studies.

Over the years a number of computational algorithms have been developed for large regression studies which attempt to find the "best" regression subset(s) \([3]\). Although these provide good results, in many cases they do not provide identical results and they cannot guarantee obtaining the "best" regression subset(s). Of these methods the most notable are a) forward-selection, b) backward-elimination, and c) stepwise regression.

Forward-selection evolves the model by successively adding the most "significant" variables, a step at a time. Backward-elimination first attempts to fit all the variables and then deletes the least "significant" variables, one step at a time. Stepwise combines forward-selection, to add variables, with backward-elimination, to remove superfluous variables. Two improved versions of stepwise, using pairwise switching, were developed by James Goodnight and are implemented under the Harris and HSWR options of the STEPWISE procedure in SAS \([2]\).

These techniques operate in a sequential manner; what happens at a given stage is highly dependent upon the results of the previous stage(s). They generally provide good models but cannot guarantee locating the "best" models for the various subset sizes.

In the late 60's and early 70's, Hocking and Leslie \([4]\) and LaMotte and Hocking \([5]\) developed a technique for isolating the one "best" regression of a size by computing only a small fraction of all regressions of that size. The following fundamental principles were employed to restrict the number of subsets to be evaluated.

If \( A \) is a set of independent variables, and \( B \) is a subset of \( A \), then \( R^2(A) > R^2(B) \).

The resulting technique would guarantee finding the one "best" regression of every size indicated and would capture in the process many good alternative regressions. These good regressions could not be guaranteed to be the "next best", although in many cases they were the "next best". This technique was a vast improvement over running all possible regressions. However, it still required great caution when applied to moderate to large studies.

In 1974 Furnival and Wilson \([6]\) reported on a "leaps and bounds" algorithm\(^*\) which could identify the "best" regression and "next best" alternative regressions for all subset sizes by computing only a small fraction of all possible

---

\(^*\) The algorithm is distributed as a set of Fortran subroutines by Professor George M. Furnival, Yale University.
2.1 Overview

The BESTREG procedure was designed to be used in concert with other SAS facilities, especially with the other regression and plotting procedures. It is intended primarily as an exploratory tool; for each regression determined it provides a summary of goodness-of-fit measures. Statistics for testing hypotheses are generally not provided; if desired, other SAS regression procedures should be used to obtain the necessary test statistics for specific regressions.

For identifying "best" regressions the user can select one of three criteria:

- RSQ - Regressions with maximum $R^2$ are determined for every subset size.
- ADJRSQ - Overall "best" regressions are determined which maximize the adjusted $R^2$, regardless of subset size.
- CP - Overall "best" regressions are determined which minimize the Mallows CP measure of total squared error, regardless of subset size.

The user can specify the number of "best" regressions desired. A "map" can be requested which shows, at a glance, the variables selected (or not selected) for all the "best" models and the interchange of variables from model to model. Specific variables can be included in every regression and the regressions can be forced through the origin. An output SAS data set can also be requested to capture various goodness-of-fit measures for all the regressions determined. Using existing plotting facilities in SAS these measures can then be readily plotted against the subset size to provide a graphical tool to assist in determining the optimal model size, the best combination of variables, and the strengths of competing models.

2.2 Specifications

The statements used for executing BESTREG are:

```
PROC BESTREG options;
  VAR variables;
  FREQ variable;
  WEIGHT variable;
  BY variables;
```

The PROC BESTREG Statement

```
PROC BESTREG options;
```

The following gives a brief description of most of the options available for BESTREG.

```
DATA=data_set_name
DATA gives the name of the SAS data set to be processed.

CRIT=selection_criterion
CRIT specifies the criterion used for identifying "best" regressions. The criteria available are:
- RSQ - maximum $R^2$ (default)
- ADJRSQ or RSQADJ - maximum adjusted $R^2$
- CP - minimum CP total squared error

These criteria are discussed in the Appendix.

```
NBEST=number
NBEST specifies the number of "best" regressions to be determined. For CRIT=RSQ, this number of "best" regressions is determined for every subset size. For ADJRSQ and CP, this number of overall "best" regressions is determined, regardless of subset size. For NBEST greater than one, the "next best" models are provided which can then be examined for possible alternatives. If NBEST is not given, one is assumed for RSQ while ten is assumed for ADJRSQ and CP.

```
NCOEF=(default=0)
NCOEF indicates the number of regressions for which to provide regression coefficients and accompanying tolerance [2] and F-ratio values.

```
OUT=data_set_name
OUT=data_set_name
An output data set can be requested to capture several goodness-of-fit measures for the "best" regressions determined. Plots of these measures versus regression size can then easily be obtained, using existing plotting facilities, to assist in determining the optimal regression size and "best" subset. The measures recorded are:
- REGNO - Regression number for cross-reference with the printed output
- P - Number of parameters estimated
- SIZE - Subset size (no. of vars. in the model)
- RSQ - $R^2$ or coefficient of determination
- ADJRSQ - $R^2$ adjusted for degree-of-freedom
- CP - Mallows' CP total squared error
- RSQADJ - Root Mean Squared Error
- F - F-ratio for the regression

FORCE=, if used, specifies the number of variables regressions can be requested and the data may be grouped.
to be forced into every regression. The variables to be forced must then appear first among
the list of variables.

\textbf{NSTART} = __ (default=1)
NSTART gives the starting regression subset size for printing, the "map", and the output data set.

\textbf{NSTOP} = __ (default=full set of predictors)
NSTOP gives the stopping regression subset size for printing, the "map", and the output data set.

\textbf{VAREST} = __
VAREST can be used to supply an external estimate of residual variance to be used in computing Cp. If not given, the mean square error (MSE) for the full model is used.

\textbf{PENALTY} = __ (default=2.0)
This option adjusts the penalty value used in the Cp calculation for variables entering the model.

\textbf{TOL} = __
The user can specify a minimum allowable tolerance [2] to be used in selecting variable subsets. The search for "best" subsets omits any subset containing an independent variable whose tolerance is less than TOL. Default value for TOL is NIND\*100/2^{11} (approx. NIND\*3.7E-7), where NIND is the number of independent variables.

\textbf{FUZZ} = __ (default=\textsc{NIND}^*100/2^{23})
Values less than FUZZ are effectively zero and thus considered as zero. The FUZZ value is used in testing for exact collinearity.

\textbf{MAP}
This option prints an incidence "map" or chart showing the variables which entered the "best" models. This "map" highlights the most frequently used variables and shows the interchange of variables between the regressions determined.

\textbf{SIGNS}
This option determines the signs of the coefficients and uses them as the markers in the "map".

\textbf{NOPRINT}
NOPRINT suppresses the printing of the regressions. The "map" and output data set are still produced if requested.

\textbf{NOINT}
NOINT forces the regressions through the origin.

\textbf{NOMAMES}
NOMAMES compacts the printed output by replacing the variable names with the variable indices.

\textbf{FULL}
If a subset of the predictor variables is collinear, the search for the "best" subsets can either be performed on the full set of predictors or can be limited to a reduced set of predictors of full rank. The FULL option forces the algorithm to search the full set of predictors. (This may cause the computational time to become excessive on moderate to large studies.)

\textbf{CORR}
CORR prints the correlation matrix.

\textbf{SSCP}
SSCP prints the sums-of-squares and cross-products matrix.

\textbf{Other SAS Information Statements}
The \textsc{VAR} statement can be given to select a subset of variables from the input data set. The \textsc{FREQ} statement can be used to process grouped data while the \textsc{WEIGHT} statement can be given to compute weighted regressions. The \textsc{BY} statement executes \textsc{BESTREG} for each \textsc{BY} group.

\textsc{BESTREG} forms the SSCP matrix for only the variables considered for the regressions. If any regression variable has a missing value, the entire observation is excluded from the analysis.

\section{3. computational time considerations}
Extreme caution must be exercised in using \textsc{BESTREG} on large studies. Computational time is affected not only by the problem size but also by the criterion selected, the number of "best" regressions desired, and the correlation structure of the data. Studies of identical size using the same selection criterion can differ widely in computational time because of differences in correlations between the variables.

To obtain an estimate of the computing times required for problems of various sizes, a number of studies were performed using randomly generated data. For the different studies, one variable was selected at random as the dependent variable and subsets of different sizes were taken as the independent variables. Timing results were recorded running both the \textsc{RSQUARE} procedure in SAS (release 82.3), generating all possible regressions, and the new \textsc{BESTREG} procedure requesting the one and ten "best" regressions of every size. (Several runs of each size were made under \textsc{BESTREG} since computational times can vary substantially.) A plot comparing the execution times of the two procedures is given by Figure 1. (Times are in minutes on an IBM 3081.)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{computation_times.png}
\caption{Comparison of computation times between \textsc{BESTREG} subset selection and \textsc{RSQUARE} all possible regressions.}
\label{fig:computation_times}
\end{figure}
It is clear why RSQUARE is not recommended for studies far beyond 15 variables. BESTREG appears to comfortably handle studies up to about 50 variables.

Further statistics on computational times are given in Table 1. Timing results are shown for numerous studies made on the 61 variable data set reported in Gibbons and McDonald [8]. It lists the computational time required for different study sizes under various parameter configurations. These results clearly show that large variations in computational times can occur under differently configured studies on the same set of data.

Table 1. Computation times of BESTREG on studies of different sizes and configurations using the 61 variable data set reported in GMR-3278.

<table>
<thead>
<tr>
<th>No. of vars.</th>
<th>best</th>
<th>Crit-</th>
<th>FULL</th>
<th>Time(min.)</th>
<th>IBM 3081</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>50</td>
<td>RSQ</td>
<td>Y</td>
<td>.008</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>RSQ</td>
<td>N</td>
<td>.007</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>10</td>
<td>RSQ</td>
<td>Y</td>
<td>.618</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>10</td>
<td>RSQ</td>
<td>N</td>
<td>.036</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>10</td>
<td>RSQ</td>
<td>N</td>
<td>.23</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>10</td>
<td>RSQ</td>
<td>Y</td>
<td>.57</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>20</td>
<td>CP</td>
<td>N</td>
<td>.26</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>20</td>
<td>CP</td>
<td>N</td>
<td>.22</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>1</td>
<td>RSQ</td>
<td>Y</td>
<td>2.33</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>10</td>
<td>RSQ</td>
<td></td>
<td>4.22</td>
<td></td>
</tr>
<tr>
<td>47(a)</td>
<td>1</td>
<td>RSQ</td>
<td>Y</td>
<td>14.17</td>
<td></td>
</tr>
<tr>
<td>47(b)</td>
<td>1</td>
<td>RSQ</td>
<td>Y</td>
<td>33.61</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>10</td>
<td>CP</td>
<td>N</td>
<td>3.07</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>CP</td>
<td>N</td>
<td>5.94</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>CP</td>
<td></td>
<td>(&gt;180)</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>10</td>
<td>RSQ</td>
<td>N</td>
<td>93.95</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>1</td>
<td>RSQ</td>
<td>N</td>
<td>193.17</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>1</td>
<td>CP</td>
<td>N</td>
<td>20.44</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>1</td>
<td>ADRSQ</td>
<td>N</td>
<td>5.04</td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>1</td>
<td>ADRSQ</td>
<td>N</td>
<td>15.64</td>
<td></td>
</tr>
</tbody>
</table>

2 On same size study, different (alpha) denote different variable sets.

3 For R² criterion, number "best" for every subset size, for ADRSQ and CP criteria, number overall "best"

4. EXAMPLES

The following gives three examples which illustrate various aspects of the BESTREG procedure.

Example #1 (contrived data)

This is a contrived simple example which demonstrates the failure of the conventional stepwise methods at identifying "best" subset regressions. Figure 2 gives the data set and the resulting output from BESTREG listing all possible regressions.

As the printout shows, variables X1 and X2 are singularly and pairwise uncorrelated with Y unless X3 is entered, in which case the triplet forms a near perfect model (R²=997). Variables X4 and X5 jointly form the "best" two variable model (R²=.924). FORWARD stepwise, even with the MAXR option [2], fails at finding the "best" three variable model since interchanges consider only one variable at a time and a double switch is required. BACKWARD stepwise identifies the

Table 1. Timing results are shown for 3 examples which illustrate possible shortcomings with using stepwise procedures.

<table>
<thead>
<tr>
<th>OBS</th>
<th>T</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.94</td>
<td>1.06</td>
<td>1.07</td>
<td>1.09</td>
<td>1.04</td>
<td>1.06</td>
</tr>
<tr>
<td>2</td>
<td>1.10</td>
<td>0.98</td>
<td>0.86</td>
<td>1.02</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>1.96</td>
<td>0.83</td>
<td>1.98</td>
<td>0.97</td>
<td>1.06</td>
<td>0.90</td>
</tr>
<tr>
<td>4</td>
<td>1.96</td>
<td>0.83</td>
<td>1.98</td>
<td>0.97</td>
<td>1.06</td>
<td>0.90</td>
</tr>
<tr>
<td>5</td>
<td>1.96</td>
<td>0.83</td>
<td>1.98</td>
<td>0.97</td>
<td>1.06</td>
<td>0.90</td>
</tr>
<tr>
<td>6</td>
<td>1.96</td>
<td>0.83</td>
<td>1.98</td>
<td>0.97</td>
<td>1.06</td>
<td>0.90</td>
</tr>
<tr>
<td>7</td>
<td>1.96</td>
<td>0.83</td>
<td>1.98</td>
<td>0.97</td>
<td>1.06</td>
<td>0.90</td>
</tr>
<tr>
<td>8</td>
<td>1.96</td>
<td>0.83</td>
<td>1.98</td>
<td>0.97</td>
<td>1.06</td>
<td>0.90</td>
</tr>
</tbody>
</table>

*Note:* Figures 1 and 2 illustrate possible shortcomings when using stepwise procedures.

---

892
"best" three variable model but for similar reasons fails to find the "best" two variable model. BESTREG finds the "best" model in all cases. Although contrived, this example should suggest the shortcomings possible with using stepwise procedures.

Example #2 (simulation study)

A set of 24 regressors was generated for 200 observations, each being normally and independently distributed as NID(0,1). Variable 25 was created as a linear combination of variables one thru five plus a random error. Variable 25 was then regressed on variables 1 thru 24, requesting the five "best" models of every size. A Cp plot was made for subsets up to size 15, as shown by Figure 3, to illustrate its utility in searching for the "best" subset in an ideal case (see Cp discussion in the Appendix). The subset size where Cp is nearest p is 5 and contains the variables 1 thru 5. (Large Cp values for the smallest subsets are off the plot.)

Example #3 (Application to a 58 variable study)

A study was made to assess the relationship between air pollution and mortality using 58 variables of the 61 variable data set reported in Gibbons and McDonald [8]. (One variable in each of the three "Home Heating Characteristics" groups was deleted since the variables in these groups formed perfect collinear subsets.) One run was made requesting the one "best" regression of every size. A second run requested the 50 "best" subsets having minimum Cp. Figure 4 is a composite plot of CP, RSD, and RMSE versus subset size. The model with Cp closest to p contained 13 variables. The minimum Cp is attained with 22 variables. (An internal GM procedure allowing multiple scales was used for plotting.)

Figures 5 and 6 present the top 50 regressions with minimum Cp. Note the dominant variables in the "map" of the 50 "best" regressions having minimum Cp and the interchange between variables 42 and 48. Also note the absence of the pollutants, variables 1-6, among the first 28 regressions. All 50 regressions are superior to the regressions of the same size obtained using the STEPWISE procedure (forward).

5. CONCLUSIONS

Unlike the conventional stepwise methods, the algorithm supporting BESTREG guarantees finding the "best" regression subset(s). The algorithm is extremely efficient and can be applied to studies including 50 or more variables. Besides reporting the one "best" regression of each size, BESTREG can produce alternative "next best" regressions with only a modest increase in cost. The graphic facilities presented condense an enormous amount of information into two visual forms which can be effectively used in identifying the optimal regression model(s).

ACKNOWLEDGMENT

The author is greatly indebted to Professor George M. Furnival, Yale University, and Robert W. Wilson, USDA, for providing the code for the "leaps and bounds" algorithm which serves as the core for the BESTREG procedure. He is further indebted to Professor Furnival for providing recent major improvements to the algorithm and for offering valuable suggestions in the development of the BESTREG procedure.

REFERENCES

APPENDIX - SELECTION CRITERIA

RSQ Criterion
The RSQ option maximizes the statistic

\[ R^2 = \frac{SST - SSE}{SST} \]

where SST and SSE are respectively the total and error sum-of-squares. Since SST is fixed, as more variables are entered, generally SSE decreases and thus \( R^2 \) increases.

ADJRSQ Criterion
The ADJRSQ option maximizes the statistic

\[ R_{a}^{2} = \frac{MST - MSE}{MST} \]

where MST and MSE are respectively the total and error mean squares. Since MST is fixed, \( R_{a}^{2} \) is maximized whenever MSE = SSE/(n-p) is minimized.

CP Criterion
The CP option minimizes the "standardized total squared error" defined as

\[ Cp = \frac{(RSS^{p}/S^2) - (n-2p)}{1} \]

where \( p \) is the number of variables in the regression including the constant if fit, \( n \) is the number of observations, RSSp is the residual sum-of-squares for the particular \( p \)-variate regression, and \( S^2 \) is an estimate of \( \sigma^2 \), the variance of the random error in the regression model.

Frequently \( S^2 \) is taken as the residual mean square resulting from the regression containing all the independent variables.

This statistic due to Mallows [7], called \( Cp \), estimates the sum of the squared biases plus the squared random errors in the response variable at all \( n \) data points. For selecting "best" subsets it is recommended that subsets be chosen in the neighborhood where both \( Cp \) is minimum and \( Cp \) equals \( p \). Subsets with small \( Cp \) are characterized by relative small estimated total squared error of prediction while subsets with \( Cp \) near \( p \) are desirable because the expected value of \( Cp \) is \( p \) when there is no bias in the fitted equation using \( p \) variables. A plot of \( Cp \) versus \( p \) is recommended as a visual aid in identifying "best" subset regressions.

---


