Selection of the Best Regression Equation by sorting out Variables

Mohammad Ehsanul Karim <wildscop@yahoo.com>
Institute of Statistical Research and training;
University of Dhaka, Dhaka – 1000, Bangladesh

The general problem to be discussed in this documentation is as follows: in many practical applications of regression analysis, the set of variables to be included in the regression model is not predetermined, and it is often the first part of the analysis to select these variables. Of course, there are some occasions when theoretical or other considerations determine the variables to be included in the equation – there the problem of variable selection does not arise. But in situations, where there is no clear-cut theory, the problem of selecting variables for a regression equation becomes an important one and this is the subject matter of our present discussion.

Suppose we have one response variable and a set of k predictor variables $X_1, X_2,...X_k$: and we wish to establish a linear regression equation for this particular response Y in terms of the basic predictor variables. We want to determine or select the best (most important or most valid) subset of the k predictors and the corresponding best-fitting regression model for describing the relationship between Y and X's. What exactly we mean by "best" depends in part on our overall goal in modeling.

Basis of Data Collection

We need to establish the basis of the data collection, as the conclusions we can make depends on this.

a) In an experiment, treatments are allocated to the experimental units and there should be elements of statistical design (such as randomization).

b) In a survey, it is not possible to allocate treatments and we take note of existing affairs. In no sense are survey data collected under controlled conditions and there may be many important factors overlooked!

Dangers of using unplanned data

When we do regression calculations by the use of any selection procedure on unplanned data arising from continuing operations and not from a designed experiment, some potentially dangerous possibilities[1] can arise: such as-

a) Errors in the model may not be random and may be due to joint effect of several variables.

b) Bias may be introduces.

c) Consequently, prediction equation becomes unreliable due to non-considering the joint interaction and confounding effects.

d) Ranges becomes invalid for prediction.

e) Large correlations between predictors are seen.

However, by randomizing out the variables, some of these problems might be avoided. Also common sense, basic knowledge of data being analyzed should be employed.

Consequences of Model Misspecification

By deleting variable from the model, we may improve the precision[2] of the parameter estimates of the retained variables even though some of the deleted variables are not negligible. This is also true for the variance of a predicted response. Deleting

[1] These problems are well described in Montgomery, Peck (1992) "Introduction to Linear Regression Analysis", 2nd ed., Page - 269.

[2] The variance of the estimates of regression coefficients for variables in the reduced equation is not greater than the variances of the corresponding estimated for the full model having all potential terms in it. Deletion of variable decrease or, more correctly, never increases the variances of estimates of the retained regression coefficients.

variables potentially introduces bias into the estimates of the coefficients of retained variables are responses. However, if the deleted variables have small effects, the MSE of the biased estimates will be less than the variable of the unbiased estimates. Thus, the amount of bias introduced is less than the reduction in the variance. There is danger in retaining negligible variables, that is, variables with zero coefficients or coefficients less than their corresponding SE from the full model. This danger is that the variances of the estimates of the parameters and the predicted response are increased.

Purposes of Regression Equations

It is not usually meaningful to speak of the 'best set' of variables to be included in the multiple regression models – there is no unique 'best set' of variables. A regression equation can be used for several purposes. The set of variables that may be best for one purpose may not be best for another. The purpose for which a regression equation is constructed should be kept in mind in the variable selection process. Some of the purposes may be broadly summarizes as follows:

Prediction: One goal is to find a model that provides best prediction of Y given $X_1, X_2, ...X_k$ for some new observation or for a batch of new observation. In practice, we emphasize estimating the regression of Y on the X's – which expresses the mean of Y as a function of the predictors. Using this goal, we may say that our best model is reliable if it predicts well in a new sample. When a regression equation is used for this purpose, the variables are selected with an eye toward minimizing the MSE of prediction.

Description: Alongside the question of prediction is the question of validity – that is, of obtaining accurate estimates for one or more regression coefficient parameters in a model and then making inferences about these parameters of interest. The purpose of the equation may be purely descriptive, to clarify the nature of complex interacting system. The goal here is to quantify the relationship between one or more independent variables of interest and dependent variable, controlling for the other variables. For this use, there are two conflicting requirements[3] – (a) to explain as much of the variation as possible (by including all or a large number of possible variables) and (b) for ease of understanding, to describe the process with as few variable as possible. In situations where description the prime goal, we try to choose the smallest number of independent variables that explains the most substantial part of the variation in the dependent variable.

Control: A regression equation may be used as a tool for control. When a regression model is used for control, accurate estimates of the parameters are important – which implies that the SE of the regression coefficients should be small. The purpose for constructing the equation may be to determine the magnitude by which the value of an independent variable must be altered to obtain a specified value of dependent variable (target response). When the regressors are highly collinear, the $\hat{b}$'s may be very poor estimates of the effects of individual regressors.

These are the broad purposes of a regression equation : occasionally which functions overlap and an equation is constructed for some or all these purposes.

[3] These conflicting objectives are well described in Montgomery, Peck (1992) "Introduction to Linear Regression Analysis", 2nd ed., Ch –7.

Steps in selecting Best Regression Model:

1. Specify the potential terms to be included in the model
2. Specify a criterion for selecting a model
3. Specify a strategy for selecting variables
4. Evaluate the model chosen

Step 1: Specify the potential terms to be included in the model

In applied fields (including the social and behavioral sciences) many independent variable (such as, future income) are not directly measurable. Under such conditions, investigators are often forced to prospect for potential independent variables[4] that could conceivably be related to the dependent variable under study.

Therefore, in the first step, we try to find the potential terms or predictor variables or functions of them that could be included in the model at any point in the process of building models. While considering these terms, we have to consider the some things such as whether there are any interaction, or multicollinearity problem, or polynomial terms or some other transformed terms to be needed. Thus, the problem of variable selection and the functional specification of the equation are linked to each other. The questions to be answered while formulating regression equation are: Which variables should be included, and in what form[5] should they be included? Although ideally the two problems (variable selection and functional specification) should be solved simultaneously, we shall for simplicity propose[6] that they be treated simultaneously: we first determine the variables that will be included in the equation, and after that, we investigate the exact form in which the variable enters. This approach is just a simplification – but it makes the problem of variable selection more tractable.

For convenience, suppose that $Z_1, Z_2, ...Z_R$, all functions of one or more of the X's, represent the complete set of variables from which the equation is to be chosen and that set includes any functions such as squares, cross-products, logarithms, inverses, and powers, thought to be desirable and necessary. Having a subset of these predictor variables (or functions of them) can create all the possible models.

However, for sake of simplicity, we discuss the models involving the simple predictor variables $X_1, X_2, ...X_k$ only for now – but the same techniques discussed below can be applied to the functional terms $Z_1, Z_2, ...Z_R$ mentioned above – and some examples are also provided in this documentation about these.

[4] After such a lengthy list has been compiled, some of the independent variables can be screened out – because of measurement error correction, duplication removing. Typically, the remaining number of independent variables that remain after the initial screening is still large.

[5] A variable can be included in the model in its original form, say x, or some transformed form, such as squares, cross-products, logarithms, inverses, powers or combination of any of such forms.

[6] This approach is proposed by Chatterjee, Price (1991) "Regression analysis by Example", 2nd ed., Ch-9.

Step 2: Specify a criterion for selecting a model

An important and crucial step in selecting the best model is to specify the selection criterion. A selection criterion is an index that can be computed for each candidate model and used to compare models. Thus, given one particular selection criterion, candidate models can be ordered from best to worst. This helps automate the process of choosing the best model. However, this selection-criterion-specific process may not find the 'best' model in a global sense. Nonetheless, using a specific selection criterion can substantially reduce the work involved in finding a 'good' model. Obviously, the selection criterion should be related to the goal of the analysis.

Many selection criterions for choosing the best model have been suggested. Two opposed criteria of selecting a resultant equation are usually involved:

1. To make the equation useful for predictive purposes, we would like our model to include as many X's as are necessary to keep bias errors small, so that reliable fitted values can be determined.
2. To keep the cost and variance of the predictions reasonably small, we should like the equation to include as few X's as possible.

The practical compromise between these extremes is what we call "Selecting the Best Regression Equation".

From statistical literatures, we present some of the criterions[7] below:

- **Criterion 1: R^2_p**

Coefficient of multiple determination, R^2_p is a ratio of sum of squares –

$$R_p^2 = \frac{RSS_p}{SSTotal} = 1 - \frac{SSE_p}{SSTotal}$$

where the denominator is constant for all possible regressions.

- Unfortunately, R^2_p provides an inadequate criterion for subset selection. Lte us explain it : R^2_p varies inversely with SSE_p, but we know that SSE can never increase as additional independent variables are included in the model. Thus R^2_p will be a maximum when all potential X variables are included in the regression equation. Thus the reason for using R^2_p cannot be to maximize R^2_p, rather the intention is to find the point where adding more X is not worthwhile because it leads to a very small increase in R^2_p. Clearly, the determination of where diminishing returns set in is a judgemental one.
- R^2_p indicates that there are p parameters in k = p-1 predictors in the regression equation on which R^2_p is based.
- R^2_p does not take account of the number of parameters in the model.
- If the predictors are randomly chosen from some distribution (say, normal) then use of R^2_p may be satisfactory. However, if the explanatory variables are fixed and controlled, then R^2_p simply reflects the controlled variation in the explanatory variables.
- For simple linear regression, with a given slope, the multiple correlation can be increased or decreased by increasing or decreasing the variation of the explanatory variables.

[7] See Wetherill, Duncombe, Kenward, Kollerstorm, Paul, Vowden (1986) "Regression Analysis with Applications" Ch- 11 for more discussion about this topic.

- **Criterion 2: R^2_{adj}**

The adjusted coefficient of multiple determination is-

$$R^2_{adj} = 1 - \frac{MESS}{\left(\frac{SSTotal}{n-1}\right)} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTotal}$$

which takes the number of parameters in the model into account through the degrees of freedom. If the constant is not included in the subset model, then (n-1) is replaced by n.

- R^2_{adj} increases if and only if MESS decreases since SSTotal/(n-1) is fixed for the given Y.
- Good or desirable models will have large values of R^2_{adj}.
- R^2_{adj} can never be negative just like R^2_p.

- **Criterion 3: C_p**

The predicted values obtained from a regression equation based on a subset of variables are generally biased. To judge the performance of an equation we should consider the MSE (with two components – variance and bias) of the predicted value rather than variance.

C.L. Mallows suggested a statistic as –

$$C_p = \frac{SSE_p}{\hat{s}^2} - (n - 2p)$$

which takes into account both the bias as well as the variance where, SSE_p is the residual sum of squares from a model containing p parameters (p is the number of parameters in the model including b_0) and $\hat{s}^2 = s^2$ is the residual mean square from the largest equation postulated containing all X's (presumed to be a reliable unbiased estimate of the error variance s^2).

- Closely related to R^2_p
- Also related to R^2_{adj}
- Has relation with partial F statistic
- $E(C_p) = p$ when there is no bias in the fitted model
- Deviation of C_p from p can be used as a measure of bias – and subsets of variables that produce values of C_p that are close to p are the desirable subsets.
- Derivation of C_p assumes unbiased estimators which may not always be true.
- One disadvantage of C_p is that it seems to be necessary to evaluate C_p for all (or most) of the possible subsets to allow interpretation. Also, sometimes, the choice may not be clear.
- Selection of good subsets are done graphically by plotting C_p versus p. Sets of variables corresponding to points close to the line C_p=p are the desirable subsets of variables to form an equation.
- Mallows suggested that good models will have negative or small C_p-p.

- **Criterion 4: Residual Mean Square (RMS)**

With a p-term equation, the RMS (or MSE) is defined as –

$$\left(RMS\right)_p = \frac{(SSE)_p}{n-p}$$

- Used for the objective of forecasting
- Between two equations, the one with the smaller RMS is usually preferred
- Related to R^2
- Also functionally related to R^2_{adj}
- Has relation with partial F statistic
- This criterion need not select the full model since both SSE_p and (n-p) will decrease as p increases; and thus RMS_p may decrease or increase. However, in most cases, this criterion will favor large subsets over smaller ones.

Step 3: Specify a strategy for selecting variables

The core step in choosing the best model is to specify the strategy for selecting variables. Such a strategy is concerned with determining how many variables and also which particular variables should be in the final model. Traditionally, such strategies have focused on deciding whether a single variable should be added to a model or whether a single variable should be deleted from a model. However, as computers became more powerful, methods for considering all models simultaneously or more than one variable per step (by generalizing single variable methods to deal with sets or chunks of variables[8]) became practical.

However, there is no unique statistical procedure or strategy for doing this[9] (then, I suppose I would not have to write this documentation this long if there was any unique procedure!). Let us consider some popular strategies:

1. All Possible Regressions Procedures
2. Backward Elimination Procedure
3. Forward Selection Procedure
4. Stepwise Regression Procedure

To add to the confusion, they do not all necessarily lead to the same solution when applied to the same problem (although for many problems, they will achieve the same answer). In fact, none of the variable selection procedures described above are guaranteed to produce the best regression equation for a given data set – this is due to the fact that there is usually not a single best equation – but rather several equally good ones.

Strategy 1: All Possible Regressions Procedures

Whenever practical, the 'all possible regressions' procedure is to be preferred over any other variable selection strategy. It alone is guaranteed to find the best model. This procedure is very direct and applied equally well to both collinear and non-linear data.

The 'all possible regressions' procedure requires that we fit each possible regression equation associated with each possible combination of the no or k independent variables that involves X_0 plus any number of the variables $X_1, X_2, ... X_k$. Since each X_i (having constant term b_0 always in the equation with or without any of these predictors, where i = 1, 2, …, k) can either be, or not be, in the equation, for k independent variables, in general, the number of models to be fitted would be $\binom{k}{0}+\binom{k}{1}+\binom{k}{2}+...+\binom{k}{k}=2^K$ (including the fit Y = b_0+e) which makes this procedure a rather cumbersome one. Once all 2^K models have been fitted, we assemble the fitted models into sets involved form 0 to k variables and then order the models within each set according to some predefined criterion. The three criteria[10] most used are:

[8] These so-called chunkwise methods for selecting variables can have subsequent advantages over single variable selection methods such as : i) they efficiently incorporate into the analysis prior knowledge and preferences about the set of variables, and ii) the number of possible models to be evaluated is reduced. For more about this Chunkwise method, readers may consult Kleinbaum, Kupper, Muller, Nizam (1998) "Applied Regression Analysis and other Multivariable methods", 3rd ed., page 400-1.

[9] If we knew the magnitude of the true random variance of the observations for any well-defined problem, our choice of a best regression equation would be much easier. Unfortunately, we are rarely in this position, so a great deal of personal judgement will be a necessary part of any of the methods discussed here.

[10] In fact, these three criteria are somewhat related to one another.

1. The value of R^2 achieved by the least squares fit.
2. The value of s^2, the residual mean square.
3. The C_p statistic.

The choice of which equation is best to use is then made by assessing the patterns observed.

When using this method, the most promising ones are identified using these criterions and then carefully analyzed by examining the residual for outliers, auto-correlation, or the need for transformations before deciding on the final model. The various subsets that are investigated may suggest interpretations of the data that might have been overlooked in a more restricted variable selection approach.

This method clearly gives an analyst the maximum amount of information available concerning the nature of relationships between Y and the set of X's. It is the only method guaranteed to find the model having the most preferred criterions (in the sense that any selection criterion will be numerically optimized for the particular sample under study).

However, naturally, using this strategy does not guarantee finding the correct (or population) model and such findings may vary from sample to sample, even though all the samples are chosen from the same population. Thus, consequently, the choice of the best model may vary from sample to sample. In fact, in many situations, several reasonable candidates for the best model can be found with different selection criteria suggesting different best models.

Also, this strategy is not always used because the amount of calculation necessary becomes 'impractical' when the number of variables that was considered in step 1 is large. While it means that the investigator has 'looked at all possibilities', it also means he has examined a large number of regression equation that intelligent thought would often reject out of hand. The amount of computer time used is wasteful and the sheer physical effort of examining all the computer printouts is enormous when more than a few variables are being examined since the number of equations and supplementary information that must be looked at may be prohibitive.

Example Data:

Let that we have the following data[11] in c:\leafbrn.dat file.

```
i     x1    x2    x3    y
1     3.05  1.45  5.67  0.34
2     4.22  1.35  4.86  0.11
3     3.34  0.26  4.19  0.38
4     3.77  0.23  4.42  0.68
5     3.52  1.10  3.17  0.18
6     3.54  0.76  2.76  0.00
7     3.74  1.59  3.81  0.08
8     3.78  0.39  3.23  0.11
9     2.92  0.39  5.44  1.53
10    3.10  0.64  6.16  0.77
11    2.86  0.82  5.48  1.17
12    2.78  0.64  4.62  1.01
13    2.22  0.85  4.49  0.89
14    2.67  0.90  5.59  1.40
15    3.12  0.92  5.86  1.05
16    3.03  0.97  6.60  1.15
17    2.45  0.18  4.51  1.49
```

[11] This data is taken from Rao (1998) "Statistical Research Methods in the life Sciences", Page 512 and available for download at http://www.stat.ncsu.edu/~st512_info/raodata/LEAFBRN.DAT

18	4.12	0.62	5.31	0.51
19	4.61	0.51	5.16	0.18
20	3.94	0.45	4.45	0.34
21	4.12	1.79	6.17	0.36
22	2.93	0.25	3.38	0.89
23	2.66	0.31	3.51	0.91
24	3.17	0.20	3.08	0.92
25	2.79	0.24	3.98	1.35
26	2.61	0.20	3.64	1.33
27	3.74	2.27	6.50	0.23
28	3.13	1.48	4.28	0.26
29	3.49	0.25	4.71	0.73
30	2.94	2.22	4.58	0.23

Getting Help From Computer Packages: MINITAB: All Possible Regressions Procedures

To get All Possible Regressions Procedures in MINITAB (version 11 was used), first we input our data as follows:

```
MTB > set c1
DATA> 3.05 4.22 3.34 3.77 3.52 3.54 3.74 3.78 2.92 3.10 2.86 2.78
2.22 2.67 3.12 3.03 2.45 4.12 4.61 3.94 4.12 2.93 2.66 3.17 2.79 2.61
3.74 3.13 3.49 2.94
DATA> end
MTB > set c2
DATA> 1.45 1.35 0.26 0.23 1.10 0.76 1.59 0.39 0.39 0.64 0.82 0.64
0.85 0.90 0.92 0.97 0.18 0.62 0.51 0.45 1.79 0.25 0.31 0.20 0.24 0.20
2.27 1.48 0.25 2.22
DATA> end
MTB > set c3
DATA> 5.67 4.86 4.19 4.42 3.17 2.76 3.81 3.23 5.44 6.16 5.48 4.62
4.49 5.59 5.86 6.60 4.51 5.31 5.16 4.45 6.17 3.38 3.51 3.08 3.98 3.64
6.50 4.28 4.71 4.58
DATA> end
MTB > name c1 'X1'
MTB > name c2 'X2'
MTB > name c3 'X3'

MTB > set c10
DATA> 0.34 0.11 0.38 0.68 0.18 0.00 0.08 0.11 1.53 0.77 1.17 1.01
0.89 1.40 1.05 1.15 1.49 0.51 0.18 0.34 0.36 0.89 0.91 0.92 1.35 1.33
0.23 0.26 0.73 0.23
DATA> end
MTB > name c10 'Y'
```

Now, we fit all possible cases using 3 predictors, which are -

Model -1: $Y \sim e$ [we usually do not show it]
Model -2: $Y \sim X_1+e$
Model -3: $Y \sim X_2+e$
Model -4: $Y \sim X_3+e$
Model -5: $Y \sim X_1+X_3+e$
Model -6: $Y \sim X_1+X_2+e$
Model -7: $Y \sim X_2+X_3+e$
Model -8: $Y \sim X_1+ X_2+X_3+e$

For convenience of separating each command from outputs, we made them bold:

```
MTB > REGRESS 'Y' on 1 predictor 'X1'        #[Model -2]

Regression Analysis


The regression equation is
Y = 2.63 - 0.592 X1

Predictor       Coef       StDev          T        P
Constant      2.6257      0.3610       7.27    0.000
X1           -0.5916      0.1085      -5.45    0.000

S = 0.3404      R-Sq = 51.5%     R-Sq(adj) = 49.8%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    1      3.4460      3.4460     29.75    0.000
Error        28      3.2435      0.1158
Total        29      6.6895

Unusual Observations
Obs        X1          Y        Fit  StDev Fit   Residual    St Resid
 19      4.61     0.1800    -0.1016     0.1572     0.2816       0.93 X

X denotes an observation whose X value gives it large influence.
```

```
MTB > REGRESS 'Y' on 1 predictor 'X2'        #[Model -3]


Regression Analysis


The regression equation is
Y = 1.01 - 0.397 X2

Predictor       Coef       StDev          T        P
Constant      1.0063      0.1303       7.72    0.000
X2           -0.3965      0.1299      -3.05    0.005

S = 0.4234      R-Sq = 25.0%     R-Sq(adj) = 22.3%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    1      1.6700      1.6700      9.32    0.005
Error        28      5.0196      0.1793
Total        29      6.6895

Unusual Observations
Obs        X2          Y        Fit  StDev Fit   Residual    St Resid
 27      2.27     0.2300     0.1061     0.2051     0.1239       0.33 X
 30      2.22     0.2300     0.1260     0.1991     0.1040       0.28 X

X denotes an observation whose X value gives it large influence.
```

```
MTB > REGRESS 'Y' on 1 predictor 'X3'        #[Model -4]

Regression Analysis


The regression equation is
Y = 0.311 + 0.0807 X3

Predictor       Coef       StDev          T        P
Constant      0.3107      0.3988       0.78    0.443
X3           0.08066     0.08360       0.96    0.343
```

```
S = 0.4809      R-Sq = 3.2%      R-Sq(adj) = 0.0%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    1      0.2152      0.2152      0.93    0.343
Error        28      6.4743      0.2312
Total        29      6.6895
```

```
MTB > REGRESS 'Y' on 2 predictor 'X1','X2'          #[Model -5]

Regression Analysis


The regression equation is
Y = 2.65 - 0.529 X1 - 0.290 X2

Predictor       Coef       StDev          T        P
Constant      2.6531      0.3157       8.40    0.000
X1          -0.52855     0.09696      -5.45    0.000
X2          -0.28996     0.09336      -3.11    0.004

S = 0.2975      R-Sq = 64.3%     R-Sq(adj) = 61.6%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    2      4.2999      2.1499     24.29    0.000
Error        27      2.3897      0.0885
Total        29      6.6895

Source       DF      Seq SS
X1            1      3.4460
X2            1      0.8538
```

```
MTB > REGRESS 'Y' on 2 predictor 'X1','X3'          #[Model -6]

Regression Analysis


The regression equation is
Y = 2.17 - 0.611 X1 + 0.111 X3

Predictor       Coef       StDev          T        P
Constant      2.1689      0.4147       5.23    0.000
X1           -0.6105      0.1037      -5.89    0.000
X3           0.11149     0.05659       1.97    0.059

S = 0.3241      R-Sq = 57.6%     R-Sq(adj) = 54.5%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    2      3.8537      1.9268     18.35    0.000
Error        27      2.8358      0.1050
Total        29      6.6895

Source       DF      Seq SS
X1            1      3.4460
X3            1      0.4077

Unusual Observations
Obs        X1          Y        Fit  StDev Fit   Residual    St Resid
```

```
 30      2.94     0.2300     0.8846     0.0687    -0.6546      -2.07R

R denotes an observation with a large standardized residual
```

MTB > REGRESS 'Y' on 2 predictor 'X2','X3'   #[Model -7]

```
Regression Analysis


The regression equation is
Y = 0.165 - 0.545 X2 + 0.206 X3

Predictor       Coef       StDev          T        P
Constant      0.1654      0.3148       0.53    0.604
X2           -0.5450      0.1268      -4.30    0.000
X3           0.20645     0.07183       2.87    0.008

S = 0.3773      R-Sq = 42.5%     R-Sq(adj) = 38.3%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    2      2.8460      1.4230     10.00    0.001
Error        27      3.8435      0.1424
Total        29      6.6895

Source       DF      Seq SS
X2            1      1.6700
X3            1      1.1761

Unusual Observations
Obs        X2          Y        Fit  StDev Fit   Residual    St Resid
 19      0.51     0.1800     0.9528     0.0928    -0.7728      -2.11R

R denotes an observation with a large standardized residual
```

MTB > REGRESS 'Y' on 3 predictor 'X1','X2','X3'   #[Model -8]

```
Regression Analysis

The regression equation is
Y = 1.81 - 0.531 X1 - 0.440 X2 + 0.209 X3

Predictor       Coef       StDev          T        P
Constant      1.8110      0.2795       6.48    0.000
X1          -0.53146     0.06958      -7.64    0.000
X2          -0.43964     0.07304      -6.02    0.000
X3           0.20898     0.04064       5.14    0.000

S = 0.2135      R-Sq = 82.3%     R-Sq(adj) = 80.2%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    3      5.5047      1.8349     40.27    0.000
Error        26      1.1848      0.0456
Total        29      6.6895

Source       DF      Seq SS
X1            1      3.4460
X2            1      0.8538
X3            1      1.2049

Unusual Observations
Obs        X1          Y        Fit  StDev Fit   Residual    St Resid
  3      3.34     0.3800     0.7973     0.0544    -0.4173      -2.02R
 10      3.10     0.7700     1.1695     0.0783    -0.3995      -2.01R
```

```
R denotes an observation with a large standardized residual
```

Now, we look at the R^2, s^2, C_p statistics to assess all the equations.

❖ Criterion-1 Check: R^2

We divide the equations in 4 sets according to inclusion of predictors (0 predictor case omitted from the table):

1 predictor	2 predictor	3 predictor
Model2 51.5%	Model5 64.3%	Model8 82.3%
Model3 25.0%	Model6 57.6%	
Model4 3.2%	Model7 42.5%	

Similarly we can find another table for R^2_{adj}

1 predictor	2 predictor	3 predictor
Model2 49.8%	Model5 61.6%	Model8 80.2%
Model3 22.3%	Model6 54.5%	
Model4 0.0%	Model7 38.3%	

Obviously, model one (with predictor X_1) has the largest value, that is, it explains the most of it! We examine the leaders in each set, and it will become clear that in 1-predictor case, model-1 and 2 (with predictor X_1 and X_2 respectively has explaining power over X_3) and in 2-predictor case, model-5 will be the preferred one (although model-6 is the closest competitor). However, in 3-predictor case, model-8 has the greatest explaining power. If we look carefully, the models with X_1 and X_2 has the most largest values over any others. So, according to this R^2 criteria, we prefer *model-5* (Note that, this is not a clear-cut solution and other information, such as knowledge of the characteristic should always be added to enable to take such decisions).

Also, examining the correlation (we have calculated this next in our discussion) of each predictor variable to dependent variable Y, we see $r_{X1,Y} = -0.71773$, $r_{X2,Y} = -0.49964$, $r_{X3,Y} = 0.17937$ where X_1 is highly correlated with Y, and X_2 is moderately correlated with Y (X_3 is correlated with Y with smaller and weaker magnitude) – which is consistent with our taken decision.

Getting Help From Computer Packages: SAS: Correlation

For correlation analysis (we calculated below more than we need now, for sake of future need), we use SAS:

```
DATA LEAFBURN;
        INFILE 'C:LEAFBRN.DAT' FIRSTOBS = 2;
        INPUT I X1 X2 X3 Y;
        X1SQ = X1*X1; X2SQ = X2*X2; X3SQ = X3*X3;
        X1X2 = X1*X2; X1X3 = X1*X3; X2X3 = X2*X3;
        LABEL   X1 = 'Percentage of Nitrogen'
                X2 = 'Percentage of Chlorine'
                X3 = 'Percentage of Potassium'
                Y  = 'Log(leafburn time) in seconds';
RUN;

PROC CORR DATA = LEAFBURN NOSIMPLE;
        TITLE 'PEARSON CORRELATION applied in leafburn data';
        VAR Y;
```

```
          WITH X1 X2 X3 X1SQ X2SQ X3SQ X1X2 X1X3 X2X3;
RUN;
```

And the output is (edited to bring down in one page):

```
PEARSON CORRELATION applied in leafburn data                1

                             Correlation Analysis

   9 'WITH' Variables:  X1       X2       X3       X1SQ     X2SQ     X3SQ
                        X1X2     X1X3     X2X3
   1 'VAR'  Variables:  Y


   Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 30

                                                          Y

                  X1                              -0.71773
                  Percentage of Nitrogen            0.0001

                  X2                              -0.49964
                  Percentage of Chlorine            0.0049


                  X3                               0.17937
                  Percentage of Potassium           0.3429

                  X1SQ                            -0.70088
                                                    0.0001

                  X2SQ                            -0.46991
                                                    0.0088

                  X3SQ                             0.16175
                                                    0.3931

                  X1X2                            -0.58094
                                                    0.0008

                  X1X3                            -0.28035
                                                    0.1335

                  X2X3                            -0.36784
                                                    0.0455
```

❖ **Criterion-2 Check: s^2**

Similarly, we can find tables with residual mean squares (s^2). First, from calculations, we find the values of s:

1 predictor	2 predictor	3 predictor
Model2 0.3404	Model5 0.2975	Model8 0.2135
Model3 0.4234	Model6 0.3241	
Model4 0.4809	Model7 0.3773	

And then square each of them, or we can get them directly from each model's ErrorMSS from the ANOVA table as follows:

1 predictor	2 predictor	3 predictor
Model2 0.1158	Model5 0.0885	Model8 0.0456
Model3 0.1793	Model6 0.1050	
Model4 0.2312	Model7 0.1424	
Average 0.1754	Average 0.1119	Average 0.0456

Sometimes, this criteria provides best cut-off point for the number of variables in regression. However, looking at the tables, we see that all ErrorMSS values are too small fraction to be compared! Thus, we choose no model according to s^2 criteria. Also, due to functional relation to s^2 with R^2, sometimes, the model chosen by these two criteria are the same.

Getting Help From Computer Packages: R: All Possible Regressions Procedures
we go for R (R.1.7.1 was used here). In R console, we give the following commands to get the output of MINITAB we just used (we do not show the outputs below):

```
> options(prompt="  R >  " )
  R >  leafbrn1<-read.table("c:\\leafbrn.dat",header=T)
  R >  leafbrn.data1<-data.frame(leafbrn1)

  R >  summary(lm(y~x1,data=leafbrn.data1));
anova(lm(y~x1,data=leafbrn.data1))
  R >  summary(lm(y~x2,data=leafbrn.data1));
anova(lm(y~x2,data=leafbrn.data1))
  R >  summary(lm(y~x3,data=leafbrn.data1));
anova(lm(y~x3,data=leafbrn.data1))
  R >  summary(lm(y~x1+x2,data=leafbrn.data1));
anova(lm(y~x1+x2,data=leafbrn.data1))
  R >  summary(lm(y~x1+x3,data=leafbrn.data1));
anova(lm(y~x1+x3,data=leafbrn.data1))
  R >  summary(lm(y~x2+x3,data=leafbrn.data1));
anova(lm(y~x2+x3,data=leafbrn.data1))
  R >  summary(lm(y~x1+x2+x3,data=leafbrn.data1));
anova(lm(y~x1+x2+x3,data=leafbrn.data1))

# However, there are other efficient ways to perform the above
analysis using R (and also S-plus) – such as –

  R >  auto.fit1 <- lm(y ~ x1, data=leafbrn.data1)
  R >  auto.fit2 <- update(auto.fit1, .~. + x2-x1)
  R >  auto.fit3 <- update(auto.fit2, .~. + x3-x2)
  R >  auto.fit4 <- update(auto.fit1, .~. + x2)
  R >  auto.fit5 <- update(auto.fit1, .~. + x3)
  R >  auto.fit6 <- update(auto.fit2, .~. +x3)
  R >  auto.fit7 <- update(auto.fit1, .~. + x2+x3) # or lm(y~.,
data=leafbrn.data1) all nominal scaled covariates are saved to the
data frame as factors.
  R >  anova(auto.fit1, auto.fit2, auto.fit3,auto.fit4,auto.fit5,
auto.fit6,auto.fit7)
Analysis of Variance Table

Model 1: y ~ x1
Model 2: y ~ x2
Model 3: y ~ x3
Model 4: y ~ x1 + x2
Model 5: y ~ x1 + x3
Model 6: y ~ x2 + x3
Model 7: y ~ x1 + x2 + x3
--( rest are omitted)--
  R >  summary(auto.fit1); summary(auto.fit2); summary(auto.fit3);
summary(auto.fit4); summary(auto.fit5); summary(auto.fit6);
summary(auto.fit7)

  R >  anova(auto.fit1); anova(auto.fit2); anova(auto.fit3);
anova(auto.fit4); anova(auto.fit5); anova(auto.fit6);
anova(auto.fit7)
```

Now, to get C_p, we command –

```
  R >  attach(leafbrn.data1)
  R >  leafbrn2 <- cbind(x1,x2,x3,y)
  R >  X<-cbind(x1,x2,x3)
  R >  Y<-y
  R >  detach()
  R >  library(leaps) # we use leaps package downloaded from CRAN
  R >  leaps(X,Y,int=TRUE, method=c("Cp"), names=NULL, df=NROW(x))

$which
      1     2     3
1  TRUE FALSE FALSE
1 FALSE  TRUE FALSE
1 FALSE FALSE  TRUE
2  TRUE  TRUE FALSE
2  TRUE FALSE  TRUE
2 FALSE  TRUE  TRUE
3  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"

$size
[1] 2 2 2 3 3 3 4

$Cp
[1]  45.17853  84.15366 116.07786  28.44100  38.23239  60.34516
4.00000
```

❖ Criterion-3 Check: C_p

We now get all the C_p values from our above calculations:

Model	C_p
2	45.17853
3	84.15366
4	116.07786
5	28.44100
6	38.23239
7	60.34516
8	4.00000

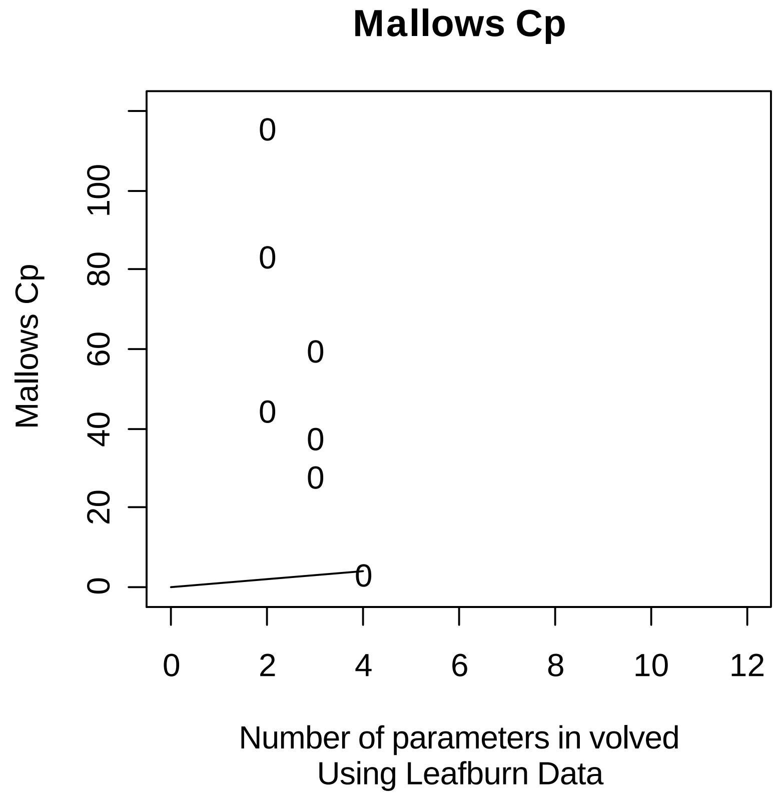
Now, we have to judge graphically. We draw the plot here: To draw the figure of (p,C_p), we write the following commands in R-console:

```
  R >  cp<-c(45.17853, 84.15366, 116.07786, 28.44100, 38.23239,
60.34516, 4.00000)
  R >  p<-c(rep(2,choose(3,1)),rep(3,choose(3,2)),rep(4,choose(3,3)))
```

# (Since, total number of models with two parameters, p are $\binom{3}{1}$ = 3, total number of models with three parameters, p are $\binom{3}{2}$ = 3, total number of models with four parameters, p are $\binom{3}{3}$ = 1)

```
  R >  plot(p,cp,pch="0", main="Mallows Cp", sub="Using Leafburn
Data", xlab="Number of parameters in volved", ylab="Mallows Cp",
xlim=c(0,12), ylim=c(0,120), type="p", axes =T, col=1)

  R >  segments(0,0,4,4) # Since full model has p = 4 and Cp = 4.
```



Note that, since we are only interested in C_p values that are close to p, in the figure largest C_p's come from models that are clearly so biased that we can eliminate them from consideration immediately. That is, regression equations with little bias will have values of C_p that fall near the line $C_p = p$. From the figure, model-1,5,6,8 are close to this line and model-8 is the closest (on the line since for the full model, $C_p = p$ exactly). But the subset model would be preferred to the full model since the full model would involve a larger total mean square error, and therefore, model 5 is the best candidate. So we choose model-5 on the basis of C_p. This matches our choice on the basis of R^2 too. However, this may not always the case.

Also, another sort of plot is recommended is p vs (C_p – p) where models with small or negative (C_p – p) are preferred. Anyway, this method also leads to the same conclusion we have just arrived.

Certain shortcuts have been suggested[12] (One of which is “Best Subset” Regression discussed next) which do not involve computing the entire set of equations while searching for the desirable subsets. But with a large number of variables, these methods still involve a considerable amount of computation.

Therefore, many alternative methods have been suggested as computationally feasible for approximating the ‘all possible regressions’ procedure. Although these methods are not guaranteed to find the best model, they can (with careful use) glean essentially all the information from the data needed to choose the best model.

“Best Subset” Regression

An alternative to performing all regressions is to use a program that provides a listing of the best E (experimenter choose E) equations with one predictor variable in, two in, three in … and so on via examining some pre-selected criterion or criterions, not by examining all 2^k equations (choice being arbitrary – so that the equations that should be included in the list, do not), b_0 being included in all these equation. Possible drawback of this procedure is, it tends to provide equations with too many predictors included.

Getting Help From Computer Packages: MINITAB: Best Subset Regression

To get “Best Subset” Regression in MINITAB (version 11 was used), we use the BREG command to get the best subset (we will go for only best 5 one here). Also note that, to allow a bit more complexity (as is the case in most of the real life data), we use the full model *including all the product and interaction terms* so that all possible cases be larger (2^9=512 possible models, where as in the previous simple case, we had 2^3=8 possible models). Also, it is worth mentioning that the following commands are the continuation of our previous MINITAB works.

```
MTB > LET c4 = (c1 * c1)
MTB > LET c5 = (c2 * c2)
MTB > LET c6 = (c3 * c3)
MTB > LET c7 = (c1 * c2)
MTB > LET c8 = (c1 * c3)
MTB > LET c9 = (c2 * c3)

MTB > name c4 'X1sq'
MTB > name c5 'X2sq'
MTB > name c6 'X3sq'
MTB > name c7 'X1X2'
MTB > name c8 'X1X3'
MTB > name c9 'X2X3'

MTB > BREG 'Y' on 9 predictor 'X1'-'X2X3';
SUBC> BEST 5.
```

[12] We call this approach as “Best subsets algorithms”. To learn more about this, readers may consult Neter, Wasserman, Kutner (1983) “Applied Linear Regression Models”, page – 428-9.

```
Best Subsets Regression #(edited)
Response is Y

                                            X X X X X X
                                            1 2 3 1 1 2
              R-Sq                    X X X s s s X X X
Vars   R-Sq   (adj)   C-p         S   1 2 3 q q q 2 3 3

   1   51.5   49.8   47.7   0.34035   X
   1   49.1   47.3   51.3   0.34864         X
   1   33.7   31.4   74.7   0.39785               X
   1   25.0   22.3   88.1   0.42340     X
   1   22.1   19.3   92.5   0.43146           X
   2   64.3   61.6   30.3   0.29750   X X
   2   63.0   60.3   32.2   0.30257   X       X
   2   62.7   60.0   32.7   0.30393     X   X
   2   62.0   59.2   33.7   0.30668   X           X
   2   61.5   58.7   34.5   0.30877         X X
   3   83.2   81.3    3.5   0.20794   X X           X
   3   82.3   80.2    4.9   0.21347   X X X
   3   82.2   80.1    5.1   0.21425     X   X       X
   3   82.0   79.9    5.4   0.21516   X X       X
   3   82.0   79.9    5.4   0.21525   X             X X
   4   85.8   83.5    1.7   0.19524   X X X       X
   4   85.6   83.3    1.9   0.19628   X X         X X
   4   85.0   82.6    2.8   0.20023   X X       X X
   4   84.7   82.3    3.2   0.20217     X X X     X
   4   84.4   81.9    3.7   0.20430     X   X     X X
   5   86.0   83.1    3.2   0.19724   X X         X X X
   5   86.0   83.1    3.3   0.19743   X X X       X   X
   5   86.0   83.1    3.3   0.19762   X X X     X X
   5   85.9   83.0    3.4   0.19791   X X X   X   X
   5   85.8   82.8    3.6   0.19902   X X X X     X
   6   86.8   83.3    4.1   0.19598   X X X   X   X   X
   6   86.4   82.9    4.6   0.19865   X X     X X X   X
   6   86.4   82.9    4.7   0.19879   X X     X   X X X
   6   86.1   82.5    5.1   0.20072   X X X   X X X
   6   86.1   82.5    5.1   0.20092   X X X       X X X
   7   86.8   82.6    6.0   0.20023   X X X X X   X   X
   7   86.8   82.6    6.1   0.20033   X X X   X X X   X
   7   86.8   82.6    6.1   0.20038   X X X   X   X X X
   7   86.7   82.5    6.2   0.20078   X X     X X X X X
   7   86.6   82.4    6.3   0.20151   X X   X X X X   X
   8   86.8   81.8    8.0   0.20472   X X X X X X X   X
   8   86.8   81.8    8.0   0.20491   X X X X X   X X X
   8   86.8   81.8    8.1   0.20499   X X X   X X X X X
   8   86.8   81.7    8.1   0.20536   X X   X X X X X X
   8   86.2   81.0    8.9   0.20933   X X X X X X X X
   9   86.8   80.9   10.0   0.20977   X X X X X X X X X
```

Now, we compare the above sets according to our criterion to choose the best ones.

Strategy 2: Backward Elimination Procedure (BE)

In the backward elimination procedure[13], we proceed as follows:

1. We determine the fitted regression equation containing all independent variables.
2. We determine the partial F statistic[14] for *every variables* in the model as thought it were the last variable to enter, and determine the partial F value or p-value associated with the test statistics.
3. Focus on the *lowest* observed F statistic (or equivalently, on the highest p-value).
4. Compare the p-value with a pre-selected significance level (called a -to-remove) or some pre-selected F value (say, F_{remove}) and decide whether to remove the variable under consideration: (a) If $F_{cal} < F_{remove}$, remove that X from consideration, that is, the variables are dropped on the basis of their smallest contribution[15] to the reduction of SSE (b) If $F_{cal} > F_{remove}$, the variable is not dropped, the backward elimination ends, and the selected model consists of variables remaining in the model.
5. Re-compute the regression equation for the remaining variables with in next iterations when one X is dropped in the previous step. We check for overall F. If significant, then we go for partial F values to hunt down the smallest one as before.

Getting Help From Computer Packages: SAS: Backward Elimination procedure

In SAS (6.12 version was used), we write the following simple program[16] to run Backward Elimination procedure:

```
DATA LEAFBURN;
        INFILE 'C:LEAFBRN.DAT' FIRSTOBS = 2;
        INPUT I X1 X2 X3 Y;
        X1SQ = X1*X1; X2SQ = X2*X2; X3SQ = X3*X3;
        X1X2 = X1*X2; X1X3 = X1*X3; X2X3 = X2*X3;
        LABEL   X1 = 'Percentage of Nitrogen'
                X2 = 'Percentage of Chlorine'
                X3 = 'Percentage of Potassium'
                Y  = 'Log(leafburn time) in seconds';
RUN;

PROC REG DATA = LEAFBURN;
        MODEL Y = X1 X2 X3 X1SQ X2SQ X3SQ X1X2 X1X3 X2X3
        /METHOD = BACKWARD SLS=0.05;
        TITLE 'BACKWARD method applied in leafburn data';
RUN;
```

And the output is (edited to bring down to continuity):

[13] This backward elimination is satisfactory, especially for statisticians who like to see all the variables in the equation once in order 'not to miss anything'.

[14] Partial F statistic test whether adding the last variable to the model significantly helps predict the dependent variable, given that the other variables are already in the model.

[15] This is equivalent to deleting the variable which has the smallest t ratio (the ratio of the regression coefficient to the standard error of the coefficient) in the equation.

[16] Note that, we used the model with product or interaction terms – not the simple one. This is due to the fact that from the simple model, no variable gets out since all are significant. So, for sake of showing the iteration, we use this large model.

```
        Backward Elimination Procedure for Dependent Variable Y

Step 0    All Variables Entered     R-square = 0.86843763   C(p) = 10.00000000

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       9             5.80943087       0.64549232      14.67   0.0001
Error           20             0.88008913       0.04400446
Total           29             6.68952000

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       2.83480862      2.31997871       0.06570163       1.49   0.2359
X1            -0.99723502      0.82571547       0.06418460       1.46   0.2412
X2            -1.48841367      0.70621440       0.19546582       4.44   0.0479
X3             0.23897966      0.67404955       0.00553140       0.13   0.7266
X1SQ           0.03272485      0.14059983       0.00238387       0.05   0.8183
X2SQ           0.15868300      0.15514366       0.04603513       1.05   0.3186
X3SQ           0.00962489      0.04916718       0.00168631       0.04   0.8468
X1X2           0.36525296      0.16678915       0.21103213       4.80   0.0406
X1X3          -0.00507684      0.14575910       0.00005338       0.00   0.9726
X2X3          -0.11113607      0.11635591       0.04014485       0.91   0.3509

Bounds on condition number:     341.5785,     13718.76
------------------------------------------------------------------------------

Step 1   Variable X1X3 Remove R-square = 0.86842965   C(p) =  8.00121315

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       8             5.80937749       0.72617219      17.33   0.0001
Error           21             0.88014251       0.04191155
Total           29             6.68952000

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       2.88166159      1.84468774       0.10227605       2.44   0.1332
X1            -1.00256670      0.79187170       0.06718164       1.60   0.2194
X2            -1.47102434      0.48745695       0.38168131       9.11   0.0066
X3             0.22012217      0.39185463       0.01322548       0.32   0.5802
X1SQ           0.03018558      0.11732818       0.00277414       0.07   0.7995
X2SQ           0.15709568      0.14472963       0.04937956       1.18   0.2900
X3SQ           0.00997938      0.04694440       0.00189397       0.05   0.8337
X1X2           0.36293282      0.14922904       0.24790168       5.91   0.0240
X2X3          -0.11249065      0.10702442       0.04630211       1.10   0.3052

Bounds on condition number:     155.8208,     6668.081
------------------------------------------------------------------------------
BACKWARD method applied in leafburn data                  4
                                                  09:15 Sunday, April 6, 1997

Step 2  Variable X3SQ Removed R-square = 0.86814652   C(p) =  6.04425360

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       7             5.80748352       0.82964050      20.69   0.0001
Error           22             0.88203648       0.04009257
Total           29             6.68952000

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       2.59480924      1.23018314       0.17837550       4.45   0.0465
X1            -0.92793959      0.69422925       0.07163042       1.79   0.1950
X2            -1.50309785      0.45334913       0.44073019      10.99   0.0031
X3             0.30117527      0.08841438       0.46521811      11.60   0.0025
X1SQ           0.01910754      0.10281339       0.00138476       0.03   0.8543
X2SQ           0.14250490      0.12462339       0.05242330       1.31   0.2651
X1X2           0.36379697      0.14590066       0.24926857       6.22   0.0207
X2X3          -0.09926809      0.08518266       0.05444786       1.36   0.2564
```

```
Bounds on condition number:     120.2678,     3217.799
------------------------------------------------------------------------------

Step 3  Variable X1SQ Removed R-square = 0.86793952   C(p) =  4.07572221

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       6             5.80609876       0.96768313      25.19   0.0001
Error           23             0.88342124       0.03840962
Total           29             6.68952000


                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       2.38083750      0.42414460       1.21024033      31.51   0.0001
X1            -0.80104107      0.12272532       1.63636896      42.60   0.0001
X2            -1.50994889      0.44226268       0.44771732      11.66   0.0024
X3             0.30357162      0.08561359       0.48292195      12.57   0.0017
X2SQ           0.14196017      0.12194598       0.05205208       1.36   0.2563
X1X2           0.36767850      0.14133487       0.25994282       6.77   0.0160
X2X3          -0.10077359      0.08299778       0.05662400       1.47   0.2370

Bounds on condition number:     73.09516,     1333.842
------------------------------------------------------------------------------

Step 4  Variable X2SQ Removed R-square = 0.86015838   C(p) =  3.25860414

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       5             5.75404668       1.15080934      29.52   0.0001
Error           24             0.93547332       0.03897805
Total           29             6.68952000

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       2.48826630      0.41703577       1.38760949      35.60   0.0001
X1            -0.79045487      0.12329024       1.60220063      41.11   0.0001
X2            -1.38346935      0.43187025       0.39999325      10.26   0.0038
X3             0.24736925      0.07122382       0.47017662      12.06   0.0020
X1X2           0.34604080      0.14114027       0.23430032       6.01   0.0219
X2X3          -0.04621282      0.06900296       0.01748274       0.45   0.5094

Bounds on condition number:     71.83096,     870.3126
------------------------------------------------------------------------------
                   BACKWARD method applied in leafburn data                  8
                                                 09:15 Sunday, April 6, 1997

Step 5 Variable X2X3 Removed  R-square = 0.85754493   C(p) =  1.65589899

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       4             5.73656394       1.43414098      37.62   0.0001
Error           25             0.95295606       0.03811824
Total           29             6.68952000


                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       2.56225921      0.39767468       1.58242492      41.51   0.0001
X1            -0.75684515      0.11136536       1.76054330      46.19   0.0001
X2            -1.45069754      0.41538339       0.46493086      12.20   0.0018
X3             0.20685617      0.03717964       1.17993845      30.95   0.0001
X1X2           0.29940802      0.12140740       0.23182986       6.08   0.0209

Bounds on condition number:     54.34849,     427.3103
BACKWARD method applied in leafburn data                  9
                                                 09:15 Sunday, April 6, 1997

------------------------------------------------------------------------------

All variables left in the model are significant at the 0.0500 level.
```

```
  Summary of Backward Elimination Procedure for Dependent Variable Y

        Variable   Number   Partial     Model
Step    Removed        In      R**2      R**2         C(p)           F  Prob>F
       Label

   1    X1X3            8    0.0000    0.8684      8.0012      0.0012  0.9726

   2    X3SQ            7    0.0003    0.8681      6.0443      0.0452  0.8337

   3    X1SQ            6    0.0002    0.8679      4.0757      0.0345  0.8543

   4    X2SQ            5    0.0078    0.8602      3.2586      1.3552  0.2563

   5    X2X3            4    0.0026    0.8575      1.6559      0.4485  0.5094
```

Strategy 3: Forward Selection Procedure (FS)

In the forward selection procedure[17], we proceed as follows:

1. The forward selection procedure starts with an equation containing no explanatory variable, only a constant term.
2. Select the first variable to enter the model the variable most highly correlated with the dependent variable, and then fit the associated straight-line regression equation.
3. If the overall F test for this regression is not significant, stop and conclude that no independent variables are important predictors. If the test is significant, include this variable in the model.
4. Determine the partial F statistic and p-value[18] associated with each remaining variable based on a regression equation containing that variable and variable initially selected.
5. Focus on the variable with the largest partial F statistic: if the test is significant (compared to predetermined F_{enter} or a -to-enter value), add the new variable to the regression equation. If not significant, we do not include the variable in our model.
6. At each subsequent step, determine the partial F statistics for the variables are not yet in the model, and then add to the model the variable with the largest partial F statistic value (if it is statistically significant). At any step, if the largest partial F statistic value is not significant, no more variables are included in the model and the process is terminated – that is, the procedure is terminated when the last variable entering the equation has an insignificant regression coefficient or all the variables are included in the equation.

Note that, forward selection procedure and backward elimination procedure does not necessarily lead to the same choice of final model. Also, it is studied that forward selection procedure tends to agree with all possible regressions for small subset sizes but not for large ones – whereas backward elimination algorithm tends to agree with all possible regressions for large subset sizes but not for small ones.

Getting Help From Computer Packages: SAS : Forward Selection procedure

In SAS (6.12 version was used), we write the following simple program (DATA step is omitted since it is just like the one used in Backward Elimination) to run Forward Selection procedure:

```
PROC REG DATA = LEAFBURN;
        MODEL Y = X1 X2 X3 X1SQ X2SQ X3SQ X1X2 X1X3 X2X3
        /METHOD = FORWARD SLE=0.10;
        TITLE 'FORWARD method applied in leafburn data';
RUN;
```

[17] Draper, Smith do not recommend forward selection procedure unless it is specially desired never to remove variables that were retained at earlier stage. For details, readers may consult Draper, Smith (1998) "Applied Regression Analysis", Third ed., Page 338-343. Also see Chatterjee, Price (1991) "Regression analysis by Example", 2nd ed., Ch-9, and Montgomery, Peck (1992) "Introduction to Linear Regression Analysis", 2nd ed., Ch –7, where they recommend the Backward elimination procedure over Forward selection procedure for variable selection. Also, backward elimination procedure is better able to handle multicollinearity than the forward selection procedure.

[18] Or some suggests that the variable that enters the equation as the second variable is one which has the highest correlation with Y given that it is adjusted for the 1st variable included.

And the output is (edited to bring down to continuity):

```
FORWARD method applied in leafburn data                  1
                                                09:11 Sunday, April 6, 1997

         Forward Selection Procedure for Dependent Variable Y

Step 1   Variable X1 Entered[19]  R-square = 0.51513526   C(p) = 47.70872455

                DF        Sum of Squares      Mean Square          F   Prob>F

Regression       1            3.44600764       3.44600764      29.75   0.0001
Error           28            3.24351236       0.11583973
Total           29            6.68952000

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       2.62570402      0.36102426       6.12740075      52.90   0.0001
X1            -0.59161367      0.10846980       3.44600764      29.75   0.0001

Bounds on condition number:            1,            1
------------------------------------------------------------------------------

                   FORWARD method applied in leafburn data                   2
                                                09:11 Sunday, April 6, 1997

Step 2   Variable X2 Entered  R-square = 0.64277444   C(p) = 30.30512594

                DF        Sum of Squares      Mean Square          F   Prob>F

Regression       2            4.29985245       2.14992623      24.29   0.0001
Error           27            2.38966755       0.08850621
Total           29            6.68952000

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       2.65313172      0.31569264       6.25118617      70.63   0.0001
X1            -0.52854935      0.09696245       2.62989015      29.71   0.0001
X2            -0.28996442      0.09335597       0.85384481       9.65   0.0044

Bounds on condition number:     1.045859,     4.183438
------------------------------------------------------------------------------

                   FORWARD method applied in leafburn data                   3
                                                09:11 Sunday, April 6, 1997

Step 3   Variable X1X3 Entered R-square = 0.83194018   C(p) =  3.54831144

                DF        Sum of Squares      Mean Square          F   Prob>F

Regression       3            5.56528044       1.85509348      42.90   0.0001
Error           26            1.12423956       0.04323998
Total           29            6.68952000

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       2.81116779      0.22258374       6.89719006     159.51   0.0001
X1            -0.83761159      0.08864063       3.86104768      89.29   0.0001
X2            -0.44384158      0.07118282       1.68109032      38.88   0.0001
X1X3           0.06396689      0.01182441       1.26542799      29.27   0.0001

Bounds on condition number:     2.102865,      15.4095
------------------------------------------------------------------------------

                   FORWARD method applied in leafburn data                   4
                                                09:11 Sunday, April 6, 1997
```

[19] X1 was checked first since it has high (-0.71773) correlation with Y. See the correlation analysis we did before.

```
Step 4   Variable X1X2 Entered  R-square = 0.85602137   C(p) =  1.88750833

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       4             5.72637209       1.43159302      37.16   0.0001
Error           25             0.96314791       0.03852592
Total           29             6.68952000

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       3.42752155      0.36741719       3.35269782      87.02   0.0001
X1            -1.01577141      0.12079559       2.72422384      70.71   0.0001
X2            -1.28629282      0.41743124       0.36581606       9.50   0.0050
X1X2           0.25065706      0.12258008       0.16109165       4.18   0.0515
X1X3           0.06178217      0.01121227       1.16974660      30.36   0.0001

Bounds on condition number:      54.8172,     434.8229
FORWARD method applied in leafburn data                  5
                                                   09:11 Sunday, April 6, 1997

------------------------------------------------------------------------------

No other variable met the 0.1000 significance level for entry into the model.

    Summary of Forward Selection Procedure for Dependent Variable Y

        Variable   Number   Partial     Model
Step    Entered        In      R**2      R**2        C(p)           F   Prob>F
       Label

   1    X1              1    0.5151    0.5151     47.7087     29.7481   0.0001
       Percentage of Nitrogen
   2    X2              2    0.1276    0.6428     30.3051      9.6473   0.0044
       Percentage of Chlorine
   3    X1X3            3    0.1892    0.8319      3.5483     29.2652   0.0001

   4    X1X2            4    0.0241    0.8560      1.8875      4.1814   0.0515
```

Strategy 4: Stepwise Regression Procedure (SW)

Stepwise regression procedure[20] is a modified version of forward selection procedure that permits reexamination, at every step, of the variables incorporated in the model in previous steps, that is, it has added proviso that at each stage the possibility of deleting a variable as in Backward elimination.

1. The stepwise regression procedure starts off by choosing an equation containing the single best X variable (here we select that X most correlated with Y and build a first order linear regression equation with that variable). We test for partial F[21] value for that variable when no X's are in the equation (b_0 is always in). If that is significant (compared to pre-selected F-to-enter value), we keep it in the model and attempts to build up with subsequent additions of X's one at a time as long as these additions are worthwhile.
2. The order of addition is determined by using some equivalent criteria[22] - (a) the largest partial F test value. (b) the largest t-statistic value. (c) the highest sample partial correlation of a variable not in the model with Y given the previously retained predictors already in the model. (d) the R^2 by judging which variable increases it more than others - to select which variable should enter next (from all the predictor variables not already in the regression model). The highest partial F value is compared to a F-to-enter (pre-selected) value.
3. After a variable has been added, the equation is examined (by testing overall regression for significance, improvement of R^2, and the partial F values for all the variables already in the model – not just the most recent entered) to see if any variable should be deleted. A variable that entered at an early stage may become superfluous at a later stage because of its relationship with other variables subsequently added to the model. To check this possibility, at each step we make a partial F test for *each variable* currently in the model, although it were the most recent variable entered, irrespective of its actual entry point into the model. The variable with the smallest (lowest) non-significant partial F value[23] (if there is such a variable) is removed and an appropriate fitted regression equation is then computed for all the remaining variables still in the model. Note that, if there is more than one non-significant partial F values, we delete only one variable with lowest non-significant partial F and we take no further action about the others. We remove only the smallest non-significant partial F variable at a time and 'think again'!
4. The model is refitted with the remaining variables: the partial Fs are obtained and similarly examined and so on. This testing of the "least useful predictor currently in the model" is carried out at every stage of the stepwise regression procedure. The whole process is continuous until no more variables can be entered (i.e., the best candidate cannot hold its place in the equation) or removed.

[20] Draper, Smith prefers and recommends stepwise regression procedure as the best of the variable selection procedures that makes economical use of computer facilities, and it avoids working with more X's than are necessary while improving the equation at every stage. For details, readers may consult Draper, Smith (1998) "Applied Regression Analysis", Third ed., Page 338 (opinion). However, Wonnacott, Wonnacott (1981) "Regression: A second course in Statistics", page 495-8 cited some problems involved in this procedure when the regressors are not orthogonal.

[21] Some programs based on stepwise regression procedure use a t-test using the relation $F(1,v)=t^2(v)$.

[22] These criteria are well-stated in Weisberg (1980) "Applied linear regression", ch – 8.

[23] Compared to an appropriate F percentage point called F-to-remove.

Choice of 'Stopping Rules' for Stepwise regression procedures

The choice of F-to-enter and F-to-remove will largely determine the character of the stepwise regression procedure. It is wise to set -

(a) F-to-enter greater than F-to-remove or
(b) α -to-enter smaller than α -to-remove

to provide 'protection' for predictors already admitted to the equation. Or else, one sometimes rejects predictors just admitted. Also, it is possible to set them both equals[24]. Another popular choice of F-to-enter and F-to-remove is equal to 4 which is roughly correspond to the 5% level of the F-distribution. On theoretical grounds, some suggest using the 25% point of the F-distribution as F_{enter} and 10% point of the appropriate F-distribution for F_{remove}[25].The choice of values for F_{enter} and F_{remove} is largely a matter of the personal preference of the analyst, and considerable latitude is often taken in this area.

Advantages of Stepwise Regression

Only variables significantly linearly related to Y are included in the model. The analyst may select minimum significance levels for inclusion or removal. Accounts for effect on whole model of adding new independent variable's rather than effect on Y in isolation. Computationally efficient.

Disadvantages of Stepwise Regression

Many t-tests (or one variable F-tests) will have been done, so there is a high probability that at least one independent variable's will have been included when it should not have been. Some multicollinearity may remain. Usually only first order terms are considered for the model, since the number of higher order terms which are potential candidates for inclusion increases rapidly with the number of independent variables. This may result in some important correlations between Y and higher order terms never being tested. Accordingly, any higher order terms that are suspected of being significantly correlated with Y should be considered in the Stepwise analysis.

A Note about Stepwise Regression

Stepwise regression can only serve as a partial tool for screening variables already being considered for inclusion. It should always be preceded by a selection process based on fundamentals & expert judgement. It should also be followed by an "All possible" analysis using the results of the stepwise, and finally by analyses in which all relevant higher order terms (polynomial & interaction etc) are included. Stepwise can again be used at this stage.

[24] Draper, Smith recommends α = 0.05 (for much conservative cases) or 0.10 for both tests. For details, readers may consult Draper, Smith (1998) "Applied Regression Analysis", Third ed., Page 342-343. Also see Montgomery, Peck (1992) "Introduction to Linear Regression Analysis", 2nd ed., page 297-8 and Wetherill, Duncombe, Kenward, Kollerstorm, Paul, Vowden (1986) "Regression Analysis with Applications" page 235-6 for related subject materials.

[25] Some authors call F-to-enter as F-IN and F-to-remove as F-OUT.

Getting Help From Computer Packages: SAS: Stepwise procedure

In SAS (6.12 version was used), we write the following simple program (DATA step is omitted since it is just like the one used in Backward Elimination) to run Stepwise procedure: here we demonstrate what happens if F-to-enter smaller than F-to-remove:

```
PROC REG DATA = LEAFBURN;
        MODEL Y = X1 X2 X3 X1SQ X2SQ X3SQ X1X2 X1X3 X2X3
        /METHOD = STEPWISE SLE=0.10 SLS = 0.05;
        TITLE 'STEPWISE method applied in leafburn data';
RUN;
```

The output is as follows:

```
STEPWISE method applied in leafburn data                 8
                                                   08:00 Sunday, April 6, 1997

              Stepwise Procedure for Dependent Variable Y

Step 1 Variable X1 Entered     R-square = 0.51513526   C(p) = 47.70872455

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       1             3.44600764       3.44600764      29.75   0.0001
Error           28             3.24351236       0.11583973
Total           29             6.68952000

                Parameter        Standard         Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       2.62570402      0.36102426       6.12740075      52.90   0.0001
X1            -0.59161367      0.10846980       3.44600764      29.75   0.0001

Bounds on condition number:            1,            1

------------------------------------------------------------------------------

STEPWISE method applied in leafburn data                  9
                                                   08:00 Sunday, April 6, 1997

Step 2   Variable X2 Entered    R-square = 0.64277444   C(p) = 30.30512594

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       2             4.29985245       2.14992623      24.29   0.0001
Error           27             2.38966755       0.08850621
Total           29             6.68952000

                Parameter        Standard         Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       2.65313172      0.31569264       6.25118617      70.63   0.0001
X1            -0.52854935      0.09696245       2.62989015      29.71   0.0001
X2            -0.28996442      0.09335597       0.85384481       9.65   0.0044

Bounds on condition number:     1.045859,     4.183438

------------------------------------------------------------------------------

STEPWISE method applied in leafburn data                 10
                                                   08:00 Sunday, April 6, 1997

Step 3   Variable X1X3 Entered  R-square = 0.83194018   C(p) =  3.54831144

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       3             5.56528044       1.85509348      42.90   0.0001
Error           26             1.12423956       0.04323998
Total           29             6.68952000
```

```
                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       2.81116779      0.22258374       6.89719006     159.51   0.0001
X1            -0.83761159      0.08864063       3.86104768      89.29   0.0001
X2            -0.44384158      0.07118282       1.68109032      38.88   0.0001
X1X3           0.06396689      0.01182441       1.26542799      29.27   0.0001

Bounds on condition number:     2.102865,      15.4095

------------------------------------------------------------------------------

STEPWISE method applied in leafburn data                1l
                                                  08:00 Sunday, April 6, 1997

Step 4    Variable X1X2 Entered R-square = 0.85602137   C(p) =  1.88750833

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       4             5.72637209       1.43159302      37.16   0.0001
Error           25             0.96314791       0.03852592
Total           29             6.68952000

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       3.42752155      0.36741719       3.35269782      87.02   0.0001
X1            -1.01577141      0.12079559       2.72422384      70.71   0.0001
X2            -1.28629282      0.41743124       0.36581606       9.50   0.0050
X1X2           0.25065706      0.12258008       0.16109165       4.18   0.0515
X1X3           0.06178217      0.01121227       1.16974660      30.36   0.0001

Bounds on condition number:      54.8172,     434.8229

------------------------------------------------------------------------------

STEPWISE method applied in leafburn data                12
                                                  08:00 Sunday, April 6, 1997
Step 5  Variable X1X2 Removed R-square = 0.83194018   C(p) =  3.54831144

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       3             5.56528044       1.85509348      42.90   0.0001
Error           26             1.12423956       0.04323998
Total           29             6.68952000

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP       2.81116779      0.22258374       6.89719006     159.51   0.0001
X1            -0.83761159      0.08864063       3.86104768      89.29   0.0001
X2            -0.44384158      0.07118282       1.68109032      38.88   0.0001
X1X3           0.06396689      0.01182441       1.26542799      29.27   0.0001
------------------------------------------------------------------------------

STEPWISE method applied in leafburn data                13
                                                  08:00 Sunday, April 6, 1997

Bounds on condition number:     2.102865,      15.4095

------------------------------------------------------------------------------

All variables left in the model are significant at the 0.0500 level.
The stepwise method terminated because the next variable to be entered was
just removed.

------------------------------------------------------------------------------
        Summary of Stepwise Procedure for Dependent Variable Y

       Variable        Number   Partial    Model
Step   Entered Removed     In      R**2     R**2      C(p)          F   Prob>F
       Label

   1   X1                   1    0.5151   0.5151   47.7087    29.7481   0.0001
       Percentage of Nitrogen
```

```
  2   X2                    2    0.1276   0.6428   30.3051     9.6473   0.0044
      Percentage of Chlorine
  3   X1X3                  3    0.1892   0.8319    3.5483    29.2652   0.0001

  4   X1X2                  4    0.0241   0.8560    1.8875     4.1814   0.0515

  5   X1X2                  3    0.0241   0.8319    3.5483     4.1814   0.0515
```

Getting Help From Computer Packages: S-plus: Stepwise procedure

To demonstrate the functionality of S-plus, we do the above procedure again in S-plus (version 4 was used):

```
> leafbrn1<-read.table("c:\\leafbrn.txt",header=T)
> leafbrn.data1<-data.frame(leafbrn1)
> attach(leafbrn.data1)
> x1sq<-x1^2;x2sq<-x2^2;x3sq<-x3^2;x1x2<-x1*x2;x1x3<-x1*x3;x2x3<-x2*x3;
> leafbrn <- cbind(leafbrn1,x1sq,x2sq,x3sq,x1x2,x1x3,x2x3)
> detach()
> leafbrn.data<-data.frame(leafbrn)
> attach(leafbrn.data)
> X<-cbind(x1,x2,x3,x1sq,x2sq,x3sq,x1x2,x1x3,x2x3) # X[26]
> Y<-y # Y[27]
> stepwise(X,Y, intercept=T, tolerance=1.e-07, method="efroymson", size.max
= ncol(X),f.crit=c(4,4), plot=T) #here F-to-enter and F-to-remove are equal

$rss:
[1] 3.2435124 2.3896675 1.1242396 0.9631479

$size:
[1] 1 2 3 4

$which[28]:
      x1 x2 x3 x1sq x2sq x3sq x1x2 x1x3 x2x3
1(+1)  T  F  F    F    F    F    F    F    F
2(+2)  T  T  F    F    F    F    F    F    F
3(+8)  T  T  F    F    F    F    F    T    F
4(+7)  T  T  F    F    F    F    T    T    F

$f.stat:
[1] 29.748064  9.647287 29.265229  4.181384

$method:
[1] "efroymson" #note that, in the process x1x2 still remains in the model
-----------------------------------------
```

A careful look at the outputs will reveal that, the results of SAS and S-plus differ (decision regarding X_1X_2) which is due to selection of SLE > SLS and f.crit values.

[26] X is a matrix of explanatory variables. It can also be a data frame. Each column represents a variable and each row represents an observation (or case). This should not contain a column of ones unless the argument intercept is FALSE. The number of rows of x should equal the length of y, and there should be fewer columns than rows. Missing values are allowed. If a data frame, x is coerced into a numeric matrix, hence factor data are transformed to numeric values using the function codes.

[27] Y is a a vector response variable. Missing values are allowed.

[28] logical matrix with as many rows as there are returned subsets. Each row is a logical vector that can be used to select the columns of x in the subset. For the forward method there are ncol(x) rows with subsets of size 1, ..., ncol(x). For the backward method there are ncol(x) rows with subsets of size ncol(x), ..., 1. For Efroymson's method there is a row for each step of the stepwise procedure. For the exhaustive search, there are nbest subsets for each size (if available). The row labels consist of the subset size with some additional information in parentheses. For the stepwise methods the extra information is +n or -n to indicate that the n-th variable has been added or dropped. For the exhaustive method, the extra information is #i where i is the subset number.

Choice of model:
While the procedures discussed do not necessarily select the absolute best model, they usually select an acceptable one. However, alternative procedures have been suggested in attempts to improve the model selection. One proposal was - Run the stepwise regression procedure with given level for acceptance and rejection. When the selection procedure stops, determine the number of variables in the final selected model. Using this number of variables, say q, do all possible sets of q variables from the k original variables and choose the best set : however, the added advantage (using stepwise regression procedure result as a priori) of this procedure is minor.

Step 4: Evaluate the model chosen

Having specified the potential terms or variables to be included in the model, the criterion for selecting a model and the strategy for applying the criterion, we must conduct the analysis as planned to get our required model. The goodness of fit of the model chosen should certainly be examined by the usual regression diagnostic methods to demonstrate that the model chosen is reasonable for the data at hand.