Please Click https://canvas.ust.hk and SFQ on the left panel To Fill out End-of-Term Course Survey.

Thanks for your attention!

Chapter 8. Variable/Model Selection

Outline

8.1. Motivation and Consequences of Variable Selection

8.1 Motivation and Consequences of Variable Selection

Introduction

In our discussion of regression problems so far we have assumed that the variables that go into the equation were **chosen in advance**. Our analysis involved examining the equation to see whether the functional specification was correct, and whether the assumptions about the error term were valid. The analysis presupposed that the set of variables to be included in the equation had already been decided. In many applications of regression analysis, however, 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. There are some occasions when theoretical or other considerations determine the variables to be included in the equation. In those situations 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.

8.1 Motivation and Consequences of Variable Selection

Introduction

The problems of **variable selection** and the **functional specification of the equation** are linked to each other. The questions to be answered while formulating a regression model are: Which variables should be included, and in what form should they be included; that is, should they enter the equation as an original variable X or as some transformed variable such as X^2, $\log X$, or a combination of both? Although ideally the two problems should be solved simultaneously, we shall for simplicity propose that they be treated sequentially. We first determine the variables that will be included in the equation, and after that investigate the exact form in which the variables enter it. This approach is a simplification, but it makes the problem of variable selection more tractable. Once the variables that are to be included in the equation have been selected, we can apply the methods described in the **earlier chapters** to arrive at the actual form of the equation.

8.1 Motivation and Consequences of Variable Selection

Formulation of Problem

We have a response variable Y and q predictor variables $X_1, X_2, \cdots, X_q$. A linear model that represents Y in terms of q variables is

$$y_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij} + \varepsilon_i, \qquad (8.1)$$

where β_j are parameters and ε_i represents random disturbances. Instead of dealing with the full set of variables (particularly when q is large), we might **delete** a number of variables and construct an equation with a subset of variables. This chapter is concerned with determining which variables are to be retained in the equation. Let us denote the set of variables **retained** by $X_1, X_2, \cdots, X_p$ and those **deleted** by $X_{p+1}, X_{p+2}, \cdots, X_q$. Let us examine the effect of variable deletion under two general conditions:

1. **The model that connects Y to the X's has all β's $(\beta_0, \beta_1, \cdots, \beta_q)$ nonzero.**
2. **The model has $\beta_0, \beta_1, \cdots, \beta_p$ nonzero, but $\beta_{p+1}, \beta_{p+2}, \cdots, \beta_q$ zero.**

Note that we are assuming the data is generated from eq. (8.1)

8.1 Motivation and Consequences of Variable Selection

Formulation of Problem

Suppose that instead of fitting (8.1), we fit the subset model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i. \tag{8.2}$$

We shall describe the effect of fitting the model to the full and partial set of X's under the two alternative situations described previously. In short, what are the effects of including variables in an equation when they should be properly left out (because the population regression coefficients are zero) and the effect of leaving out variables when they should be included (because the population regression coefficients are not zero)? We will examine the effect of deletion of variables on the estimates of parameters and the predicted values of Y. The solution to the problem of variable selection becomes a little clearer once the effects of retaining unessential variables or the deletion of essential variables in an equation are known.

8.1 Motivation and Consequences of Variable Selection

Compared to Model Testing (Chapter 3)

In Chapter 3, we have designed a method for testing

$$H_0(\textit{reduced model}) : Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$
$$H_1(\textit{full model}) : Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \cdots + \beta_q X_q$$

The test statistic is

$$F = \frac{[\text{SSE}(RM) - \text{SSE}(FM)]/(q-p)}{\text{SSE}(FM)/(n-q-1)}$$

Under the null hypothesis (that is, assuming the reduced model is true), we have this F-statistic follows an F distribution with degrees of freedom $q-p$ and $n-q-1$.

Basically, this tells us that if the reduced model is true, what happens for this F-statistic.

8.1 Motivation and Consequences of Variable Selection

Compared to Model Testing (Chapter 3)

However, in this section, let us consider the following question:

Suppose that the true model (i.e., the way data is generated) is the **full model**, but we actually use the **reduced model** to fit the data, then what happens to the estimated coefficients?

This is different from the F-statistic (last slide), whose distribution under null hypothesis is based on the assumption that **the true model is the reduced model**.

8.1 Motivation and Consequences of Variable Selection

Consequences of Variable Selection

$$y_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij} + \varepsilon_i, \tag{8.1}$$

On the bias and variance

Denote the estimates of the regression parameters by $\hat{\beta}_0^*, \hat{\beta}_1^*, \cdots, \hat{\beta}_q^*$ when the model (8.1) is fitted to the full set of variables $X_1, X_2, \cdots, X_q$. Denote the estimates of the regression parameters by $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ when the model (8.2) is fitted. Let $\hat{y}_i^*$ and $\hat{y}_i$ be the **predicted values** from the **full** and **partial set** of variables corresponding to an observation $(x_{i1}, x_{i2}, \cdots, x_{iq})$. The results can now be summarized as follows: $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ are **biased** estimates of $\beta_0, \beta_1, \cdots, \beta_p$ unless the remaining β's in the model $(\beta_{p+1}, \beta_{p+2}, \cdots, \beta_q)$ are **zero** or the variables $X_1, X_2, \cdots, X_p$ are orthogonal to the variable set $(X_{P+1}, X_{p+2}, \cdots, X_q)$. The estimates $\hat{\beta}_0^*, \hat{\beta}_1^*, \cdots, \hat{\beta}_p^*$ have **less precision** than $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ that is,

$$\mathrm{Var}(\hat{\beta}_j^*) \geq \mathrm{Var}(\hat{\beta}_j), \quad j = 0, 1, \cdots, p.$$

Note that we are assuming the data is generated by model (8.1)

Result

The variance of the estimates of regression coefficients for variables in the **reduced equation** are not greater than the variances of the corresponding estimates for the **full model**. **Deletion of variables decreases or, more correctly, never increases the variances of estimates of the retained regression coefficients.**

8.1 Motivation and Consequences of Variable Selection

Theoretical Understanding

Data modelling

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} & x_{1(p+1)} & \cdots & x_{1q} \\ 1 & x_{21} & \cdots & x_{2p} & x_{2(p+1)} & \cdots & x_{2q} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} & x_{n(p+1)} & \cdots & x_{nq} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \\ \beta_{p+1} \\ \vdots \\ \beta_q \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$\mathbf{X}_1$

Data generated by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Using all the variables for estimation

$$\begin{aligned} \hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \end{aligned}$$

The bias $E(\hat{\boldsymbol{\beta}}^*) - \boldsymbol{\beta} = \mathbf{0}$, so $\hat{\boldsymbol{\beta}}^*$ is **unbiased**.

The covariance $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) = \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$

8.1 Motivation and Consequences of Variable Selection

Theoretical Understanding

Note that we are assuming the data is generated by model (8.1)

Using the partial set of variables for estimation

$$\begin{aligned}\hat{\boldsymbol{\beta}} =& (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \\ =& (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{\varepsilon}\end{aligned}$$

Its bias equals $E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}$ which is not zero unless $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$.

Its covariance matrix

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \cdot (\mathbf{X}_1'\mathbf{X}_1)^{-1}$$

In contrast to the covariance $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) = \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$.

8.1 Motivation and Consequences of Variable Selection

Consequences of Variable Selection

Definition of Mean Squared Error (MSE)

The mean squared error of an estimator $\hat{\theta}$ to estimate the parameter θ is defined by

$$\text{MSE}(\hat{\theta}) = \underbrace{\text{Var}(\hat{\theta})}_{\text{Variance}} + \underbrace{(\mathbb{E}\hat{\theta} - \theta)^2}_{\text{Bias}^2}$$

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

Note that we are assuming the data is generated by model (8.1)

Comparison with reduced regression estimators

Here we are considering the general case that $\beta_{p+1}, \beta_{p+2}, \cdots, \beta_q$ may be non-zero.

Since $\hat{\beta}_j$ are biased and $\hat{\beta}_j^*$ are not, a better comparison of the precision of estimates would be obtained by comparing the mean square errors of $\hat{\beta}_j$ with the variances of $\hat{\beta}_j^*$. The mean squared errors (MSE) of $\hat{\beta}_j$ will be smaller than the variances of $\hat{\beta}_j^*$ only if the deleted variables have regression coefficients smaller in magnitude than the standard deviations of the estimates of the corresponding coefficients. The estimate of σ^2, based on the subset model, is generally **biased upward**.

Since $\hat{\beta}_j^*$ is unbiased, the $\text{MSE}(\hat{\beta}_j^*) = \text{Var}(\hat{\beta}_j^*)$.

Since we have deleted some variables, the data may be fitted not as well as the full model.

8.1 Motivation and Consequences of Variable Selection

Consequences of Variable Selection

Note that we are assuming the data is generated by model (8.1)

On the prediction

Let us now look at the effect of deletion of variables on prediction. The prediction $\hat{y}_i$ is biased unless the deleted variables have zero regression coefficients, or the set of retained variables are orthogonal to the set of deleted variables. The variance of a predicted value from the subset model is smaller than or equal to the variance of the predicted value from the full model; that is,

$$\text{Var}(\hat{y}_i) \leq \text{Var}(\hat{y}_i^*).$$

The conditions for $\text{MSE}(\hat{y}_i)$ to be smaller than $\text{Var}(\hat{y}_i^*)$ are identical to the conditions for $\text{MSE}(\hat{\beta}_j)$ to be smaller than $\text{Var}(\hat{\beta}_j^*)$, which we have already stated.

$$\hat{y}_i = (1, x_{i1}, \cdots, x_{ip})\hat{\boldsymbol{\beta}} \quad \text{and} \quad \hat{y}_i^* = (1, x_{i1}, \cdots, x_{ip}, x_{i(p+1)}, \cdots, x_{iq})\hat{\boldsymbol{\beta}}^*$$

using only the partial set of predictors

using all the predictors

8.1 Motivation and Consequences of Variable Selection

Consequences of Variable Selection

Note that we are assuming the data is generated by model (8.1)

The bias of $\hat{y}_i$ is defined by $E(\hat{y}_i) - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq})$ and its variance is obtained by

$$\mathrm{Var}(\hat{y}_i) = (1, x_{i1}, \cdots, x_{ip}) \cdot \mathrm{Cov}(\hat{\boldsymbol{\beta}}) \cdot \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$$

Then the mean squared error of $\hat{y}_i$ is given by

$$\mathrm{MSE}(\hat{y}_i) = \big(E(\hat{y}_i) - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq})\big)^2 + \mathrm{Var}(\hat{y}_i)$$

The bias of $\hat{y}_i^*$ is defined by $E(\hat{y}_i^*) - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}) = 0$ and its variance is obtained by

$$\mathrm{Var}(\hat{y}_i^*) = (1, x_{i1}, \cdots, x_{iq}) \cdot \mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) \cdot \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{iq} \end{pmatrix}$$

Then the mean squared error of $\hat{y}_i^*$ is given by

$$\mathrm{MSE}(\hat{y}_i^*) = \mathrm{Var}(\hat{y}_i^*)$$

8.1 Motivation and Consequences of Variable Selection

Consequences of Variable Selection

Why deleting some variables benefits the MSE and prediction?

The rationale for variable selection can be outlined as follows: Even though the variables deleted have nonzero regression coefficients, the regression coefficients of the retained variables may be estimated with **smaller variance** from the subset model than from the full model. The same result also holds for the variance of a predicted response. The price paid for deleting variables is in the **introduction of bias** in the estimates. However, there are conditions (as we have described above), when the MSE of the biased estimates will be smaller than the variance of their unbiased estimates; that is, the gain in precision is **not offset** by the square of the bias. On the other hand, if some of the retained variables are extraneous or unessential, that is, have zero coefficients or coefficients whose magnitudes are smaller than the standard deviation of the estimates, the inclusion of these variables in the equation leads to a loss of precision in estimation and prediction.

8.2. Uses of Regression Equations

8.2 Uses of Regression Equations

A regression equation has many uses. These are broadly summarized below.

Description and Model Building

A regression equation may be used to describe a given process or as a model for a complex interacting system. The purpose of the equation may be purely descriptive, to clarify the **nature of this complex interaction**. For this use there are two **conflicting** requirements: (1) to account for as much of the variation as possible, which points in the direction for inclusion of a large number of variables; and (2) to adhere to the principle of **parsimony**, which suggests that we try, for ease of understanding and interpretation, to describe the process with as few variables as possible. In situations where description is the prime goal, we try to choose the smallest number of predictor variables that accounts for the most substantial part of the variation in the response variable.

Estimation and Prediction

A regression equation is sometimes constructed for **prediction**. From the regression equation we want to predict the value of a future observation or estimate the mean response corresponding to a given observation. When a regression equation is used for this purpose, the variables are selected with an eye toward minimizing the MSE of prediction.

Control

A regression equation may be used as a tool for control. The purpose for constructing the equation may be to determine the magnitude by which the value of a predictor variable must be altered to obtain a specified value of the response (target) variable. Here the regression equation is viewed as a response function, with Y as the response variable. For instance, how much more ads are needed to boost the revenue to 1 million dollars? For control purposes it is desired that the coefficients of the variables in the equation be measured accurately; that is, the standard errors of the regression coefficients are small.

8.2 Uses of Regression Equations

These are the broad uses of a regression equation. Occasionally, these functions overlap and an equation is constructed for some or all of these purposes. The main point to be noted is that the purpose for which the regression equation is constructed determines the criterion that is to be optimized in its formulation. It follows that a subset of variables that may be best for one purpose may not be best for another. The concept of the "best" subset of variables to be included in an equation always requires additional qualification.

Before discussing actual selection procedures we make two preliminary remarks. First, it is not usually meaningful to speak of the "best set" of variables to be included in a multiple regression equation. 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. We shall show later that the purpose for which an equation is constructed determines the criteria for selecting and evaluating the contributions of different variables.

8.2 Uses of Regression Equations

Second, since there is no best set of variables, there may be several subsets that are adequate and could be used in forming an equation. A good variable selection procedure should point out these several sets rather than generate a so-called **single** "best" set. The various sets of adequate variables throw light on the structure of data and help us in understanding the underlying process. In fact, the process of variable selection should be viewed as an intensive analysis of the correlational structure of the predictor variables and how they individually and jointly affect the response variable under study. These two points influence the methodology that we present in connection with variable selection.

8.3. Criteria for Evaluating Equations

Class Discussion

For two model equations with the same number of predictors, e.g.,

$$\text{Model 1} : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

versus

$$\text{Model 2} : Y = \alpha_0 + \alpha_1 X_3 + \alpha_2 X_4 + \varepsilon$$

Given a data set of n observations $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$, how do we compare the two models, i.e., which one is preferable?

Can we use the general framework introduced in Chapter 3 for model testing, i.e., reduced model versus full model ?

Class Discussion

For two model equations with different numbers of predictors, e.g.,

$$\text{Model 1}: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

versus

$$\text{Model 2}: Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_3 + \alpha_3 X_4 + \varepsilon$$

Given a data set of n observations $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$, how do we compare the two models, i.e., which one is preferable?

Can we use the general framework introduced in Chapter 3 for model testing, i.e., reduced model versus full model ?

By Chapter 3, for a model with $p-1$ predictors and 1 intercept, we have

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \quad \text{and} \quad R_a^2 = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)}$$

8.3 Criteria for Evaluating Equations

Residual Mean Square

To judge the adequacy of various fitted equations we need a criterion. Several have been proposed in the statistical literature. We describe the two that we consider most useful.

Residual Mean Square

One measure that is used to judge the **adequacy of a fitted equation** is the residual mean square (RMS). With a p-term equation (includes a constant and $p-1$ variables), the RMS is defined as

$$\mathrm{RMS}_p = \frac{\mathrm{SSE}_p}{n-p}. \tag{8.3}$$

where SSE_p is the residual sum of squares for a p-term equation. Between two equations, the one with the smaller RMS is usually preferred, especially if the objective is forecasting.

Here we put a subscript p to emphasize that it is for a model with p terms

It is clear that RMS_p is related to the square of the multiple correlation coefficient R_p^2 and the square of the adjusted multiple correlation coefficient R_{ap}^2 which have already been described (Chapter 3) as measures for judging the adequacy of fit of an equation. Here we have added a subscript to R^2 and R_a^2 to denote their dependence on the number of terms in an equation. The relationship between these quantities are given by

$$R_p^2 = 1 - (n-p)\frac{\mathrm{RMS}_p}{(\mathrm{SST})} \tag{8.4}$$

and

$$R_{ap}^2 = 1 - (n-1)\frac{\mathrm{RMS}_p}{(\mathrm{SST})}, \tag{8.5}$$

same to Chapter 3

where

$$\mathrm{SST} = \sum (y_i - \bar{y})^2.$$

8.3 Criteria for Evaluating Equations

Mallows C_p

Residual Mean Square

Note that R^2_{ap} is more appropriate than R^2_p when comparing models with different number of predictors because R^2_{ap} adjusts (penalizes) for the number of predictor variables in the model.

The model with larger adjusted R^2 is preferred

Mallows C_p

We pointed out earlier that 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 mean square error of the predicted value rather than the variance. The standardized total mean squared error of prediction for the observed data is measured by

$$J_p = \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathrm{MSE}(\hat{y}_i), \tag{8.6}$$

where $\mathrm{MSE}(\hat{y}_i)$ is the mean squared error of the ith predicted value from a p-term equation, and σ^2 is the variance of the random errors. The $\mathrm{MSE}(\hat{y}_i)$ has two components, the variance of prediction arising from estimation and a bias component arising from the deletion of variables.

8.3 Criteria for Evaluating Equations

Mallows C_p

Mallows C_p

To estimate J_p, Mallows (1973) uses the statistic

Based on Chapter 3, when the model is correct, $E[\text{SSE}_p] = (n-p)\sigma^2$

$$C_p = \frac{\text{SSE}_p}{\hat{\sigma}^2} + (2p - n), \tag{8.7}$$

where $\hat{\sigma}^2$ is an estimate of σ^2 and is usually obtained from the linear model with the full set of q variables. It can be shown that the expected value of C_p is p when there is no bias in the fitted equation containing p terms. Consequently, the deviation of C_p from p can be used as a measure of bias. The C_p statistic therefore measures the performance of the variables in terms of the standardized total mean square error of prediction for the observed data points irrespective of the unknown true model. It takes into account both the bias and the variance. Subsets of variables that produce values of C_p that are close to p are the desirable subsets. The selection of "good" subsets is done graphically. For the various subsets a graph of C_p is plotted against p. The line $C_p = p$ is also drawn on the graph. Sets of variables corresponding to points close to the line $C_p = p$ are the good or desirable subsets of variables to form an equation. The use of C_p plots is illustrated and discussed in more detail in the example that is given in Section 8.6.

To use Mallows' C_p, we must designate a full model to obtain $\hat{\sigma}^2$.

Example

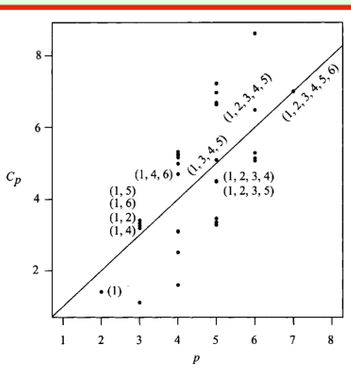

Figure 11.1 Supervisor's Performance data: Scatter plot of C_p versus p for subsets with $C_p < 10$.

8.3 Criteria for Evaluating Equations

Information Criteria

Akaike Information Criterion (AIC)

Variable selection in the regression context can be viewed as a model selection problem. The Information criteria that we now describe arose first in the general problem of model selection. The Akaike (1973) Information Criterion (AIC) in selecting a model tries to balance the **conflicting** demands of accuracy (fit) and simplicity (small number of variables). This is the principle of parsimony already discussed in Chapter 3. AIC for a p-term equation (a constant, and $p-1$ variables) is given by

$$\text{AIC}_p = n \ln(\text{SSE}_p/n) + 2p. \qquad (8.8)$$

The models with smaller AIC are preferred. **The model with smaller AIC is preferred**

We can see from (8.8) that for two models with similar SSE, AIC penalIzes the model that has a larger number of variables. The numerical value of AIC for a single model is not very meaningful or descriptive. AIC can be used, however, to rank the models on the basis of their twin criteria of fit and simplicity. Models with AIC not differing by 2 should be treated as equally adequate. Larger differences in AIC indicate significant difference between the quality of the models. The one with the lower AIC should be adopted.

8.3 Criteria for Evaluating Equations

Information Criteria

Akaike Information Criterion (AIC)

A great advantage of AIC is that it allows us to compare **non-nested** models. A group of models are **nested** if they can be obtained from a larger model as special cases (see Chapter 3). We cannot perform an F-Test, for example, to compare the adequacy of a model based on (X_1, X_2, X_3) with one based on (X_4, X_5). The choice of these two sets of variables may be dictated by the nature of the problem at hand. The AIC will allow us to make such comparisons but not the F-Test described earlier.

To compare models by AIC we must have complete data (no missing values). The AIC must be calculated on the same set of observations. If there are many missing values for some variables, application of AIC may be inefficient because observations in which some variables were missing will be dropped.

Bayes Information Criterion (AIC)

The model with smaller BIC is preferred

Several modifications of AIC have been suggested. One popular variation called Bayes Information Criterion (BIC), originally proposed by Schwarz (1978), is defined as

$$\text{BIC}_p = n \ln(\text{SSE}_p/n) + p(\ln n). \tag{8.9}$$

The difference between AIC and BIC is in the severity of penalty for p. The penalty is far more severe in BIC when $n > 8$. This tends to control the overfitting (resulting in a choice of larger p) tendency of AIC.

8.3 Criteria for Evaluating Equations

Example on Supervisor Performance Data

Table 3.2 Description of Variables in Supervisor Performance Data

Variable	Description
Y	Overall rating of job being done by supervisor
X_1	Handles employee complaints
X_2	Does not allow special privileges
X_3	Opportunity to learn new things
X_4	Raises based on performance
X_5	Too critical of poor performance
X_6	Rate of advancing to better jobs

Model 1: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_4 + \varepsilon$

Model 2: $Y = \alpha_0 + \alpha_1 X_3 + \alpha_2 X_4 + \alpha_3 X_5 + \varepsilon$

Table 3.3 Supervisor Performance Data

Row	Y	X_1	X_2	X_3	X_4	X_5	X_6
1	43	51	30	39	61	92	45
2	63	64	51	54	63	73	47
3	71	70	68	69	76	86	48
4	61	63	45	47	54	84	35
5	81	78	56	66	71	83	47
6	43	55	49	44	54	49	34
7	58	67	42	56	66	68	35
8	71	75	50	55	70	66	41
9	72	82	72	67	71	83	31
10	67	61	45	47	62	80	41
11	64	53	53	58	58	67	34
12	67	60	47	39	59	74	41
13	69	62	57	42	55	63	25
14	68	83	83	45	59	77	35
15	77	77	54	72	79	77	46
16	81	90	50	72	60	54	36
17	74	85	64	69	79	79	63
18	65	60	65	75	55	80	60
19	65	70	46	57	75	85	46
20	50	58	68	54	64	78	52
21	50	40	33	34	43	64	33
22	64	61	52	62	66	80	41
23	53	66	52	50	63	80	37
24	40	37	42	58	50	57	49
25	63	54	42	48	66	75	33
26	66	77	66	63	88	76	72
27	78	75	58	74	80	78	49
28	48	57	44	45	51	83	38
29	85	85	71	71	77	74	55
30	82	82	39	59	64	78	39

8.3 Criteria for Evaluating Equations

Example on Supervisor Performance Data

By adjusted R-squared

```
> ################ Example on Supervisor Performance dataset
> Sup_Prf <- read.table("data/P060.txt",header=TRUE)  ## read data
> mod1<-lm(Y~X1+X4,data=Sup_Prf)    ## model 1
> mod2<-lm(Y~X3+X4+X5,data=Sup_Prf)    ## model 2
>
> ###### use adjusted R squared
> crit<-data.frame(model1=rep(0,4),model2=rep(0,4))
> row.names(crit)<-c("Adjusted R^2","Mallow's Cp","AIC","BIC")
> crit[1,]<-c(summary(mod1)$adj.r.squared, summary(mod2)$adj.r.squared)
> crit[1,]
                model1    model2
Adjusted R^2 0.660483 0.3875397
```

8.3 Criteria for Evaluating Equations

Example on Supervisor Performance Data

By Mallow's Cp

```
> ##### Mallow's Cp by direct R function
> library(olsrr)
> crit[2,]<-c(ols_mallows_cp(mod1,mod2),ols_mallows_cp(mod2,mod2))    ## This is a run use of Mallows' Cp sicne model2 is not the full model
> crit[2,]
               model1 model2
Mallow's Cp -9.032565      4
```

The above R codes represents an incorrect use of Mallows' C_p since model 2 is not the full model when we compare model 1 and model 2. For this purpose, we can simply introduce a full model so that model 1 and model 2 are both its reduced model. An example is

$$\text{Model 3}: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon$$

It can be viewed as full model, and the *Model 1* and *Model 2* can be both viewed as the reduced model.

Here, you can also choose the full model as including all the 8 predictors.

```
> mod3<-lm(Y~X1+X3+X4+X5,data=Sup_Prf)    ## introduce model 3, which can be viewed as the full model for model1 and model2
> crit[2,]<-c(ols_mallows_cp(mod1,mod3),ols_mallows_cp(mod2,mod3))
>
> crit[c(1,2),]
                 model1     model2
Adjusted R^2 0.660483  0.3875397
Mallow's Cp  3.095767 25.0681808
```

8.3 Criteria for Evaluating Equations

Example on Supervisor Performance Data

By AIC

```
> ######  AIC by direct R function
> library(stats)
> crit[3,]<-c(AIC(mod1),AIC(mod2))
> crit[c(1,2,3),]
                    model1       model2
Adjusted R^2      0.660483    0.3875397
Mallow's Cp       3.095767   25.0681808
AIC             207.519625 226.0862160
```

8.3 Criteria for Evaluating Equations

Example on Supervisor Performance Data

By BIC

```
> ######  BIC by direct R function
> library(stats)
> crit[4,]<-c(BIC(mod1),BIC(mod2))
> crit
                   model1      model2
Adjusted R^2     0.660483   0.3875397
Mallow's Cp      3.095767  25.0681808
AIC            207.519625 226.0862160
BIC            213.124414 233.0922029
```

8.4. Collinearity and Variable Selection

8.4 Collinearity and Variable Selection

Two Situations

In discussing variable selection procedures, we distinguish between two broad situations:

1. The predictor variables are not collinear; that is, there is no strong evidence of collinearity.

2. The predictor variables are collinear; that is, the data are highly multicollinear.
not covered in this course

Depending on the correlation structure of the predictor variables, we propose different approaches to the variable selection procedure. If the data analyzed are not collinear, we proceed in one manner, and if collinear, we proceed in another.

As a first step in variable selection procedure we recommend calculating the **variance inflation factors** (VIFs) or the **eigenvalues of the correlation matrix** of the predictor variables. If none of the VIFs are greater than 10, collinearity is not a problem. Further, as we explained in Chapter 7, the presence of small eigenvalues indicates collinearity. If the condition number is larger than 15, the variables are collinear. We may also look at the sum of the reciprocals of the eigenvalues. If any of the individual eigenvalues are less than 0.01, or the sum of the reciprocals of the eigenvalues is greater than, say, five times the number of predictor variables in the problem, we say that the variables are collinear. If the conditions above do not hold, the variables are regarded as noncollinear.

The condition number is defined by $\kappa = \sqrt{\lambda_{\max}/\lambda_{\min}}$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum eigenvalues of the matrix of correlation coefficients.

8.4 Collinearity and Variable Selection

Variable Selection by Evaluating All Possible Equations

The first procedure described is very **direct** and applies equally well to both **collinear** and **non-collinear** data. The procedure involves fitting **all** possible subset equations to a given body of data. With q variables the total number of equations fitted is 2^q (including an equation that contains all the variables and another that contains no variables). The latter is simply $\hat{y}_i = \bar{y}$, which is obtained from fitting the model $Y = \beta_0 + \varepsilon$. 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.

However, the number of equations and supplementary information that must be looked at maybe prohibitively large. Even with only six predictor variables, there are $64(2^6)$ equations to consider; with seven variables the number grows to $128(2^7)$, neither feasible nor practical. An efficient way of using the results from all possible equations is to pick out the three "best" (on the basis of R^2, C_p , RMS, or the information criteria outlined earlier) equations containing a specified number of variables. This **smaller subset** of equations is then analyzed to arrive at the final model. These regressions are then carefully analyzed by examining the residuals for outliers, collinearity, 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.

8.4 Collinearity and Variable Selection

Variable Selection by Evaluating All Possible Equations

When the number of variables is large, the evaluation of all possible equations may not be practically feasible. Certain shortcuts have been suggested 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. There are variable selection procedures that do not require the evaluation of all possible equations. Employing these procedures will not provide the analyst with as much information as the fitting of all possible equations, but it will entail considerably less computation and may be the only available practical solution. These are discussed in Section 8.5. These procedures are quite efficient with noncollinear data. **We do not, however, recommend them for collinear data.**

8.5. Variable Selection Procedures

8.5 Variable Selection Procedures

For cases when there are a large number of potential predictor variables, a set of procedures that does not involve computing of all possible equations has been proposed. These procedures have the feature that the variables are **introduced** or **deleted** from the equation **one at a time**, and involve examining only a subset of all possible equations. With q variables these procedures will involve evaluation of at most $q+1$ equations, as contrasted with the evaluation of 2^q equations necessary for examining all possible equations. The procedures can be classified into two broad categories: (1) the **forward selection** (FS) procedure, and (2) the **backward elimination** (BE) procedure. There is also a very popular modification of the FS procedure called the stepwise method. The three procedures are described and compared below.

8.5 Variable Selection Procedures

Forward Selection Procedure

The forward selection procedure starts with an equation containing no predictor variables, only a constant term. The **first** variable included in the equation is the one which has the highest simple correlation with the response variable Y. If the regression coefficient of this variable is **significantly different** from zero it is retained in the equation, and a search for a **second variable** is made. The variable that enters the equation as the second variable is one which has the highest correlation with Y, after Y has been adjusted for the effect of the first variable, that is, the variable with the highest simple correlation coefficient with the **residuals** from Step 1.

The significance of the regression coefficient of the **second variable** is then tested. If the regression coefficient is **significant**, a search for a **third variable** is made in the same way. 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. The significance of the regression coefficient of the last variable introduced in the equation is judged by the standard t-Test computed from the latest equation. Most forward selection algorithms use a low t cutoff value for testing the coefficient of the newly entered variable; consequently, the forward selection procedure goes through the full set of variables and provides us with $q + 1$ possible equations.

Then, among the q+1 equations, which one should we choose?

8.5 Variable Selection Procedures

Forward Selection Procedure

(a) **Step 1**: From all the variables $X_1, \cdots, X_q$, choose the one with the **hightest** correlation with Y, *say it is* X_1. Then, we get the model: $Y = \beta_0 + \beta_1 X_1$. Fit the model, and test against $\beta_1 = 0$, if it is significant (by t-value), then retain X_1 in the model and compute the residual $\mathbf{e}$; if it is insignificant, stop the whole procedure.

(b) **step 2**: From the remaining variables $X_2, \cdots, X_q$, choose the one with the **highest** correlation with $\mathbf{e}$, *say it is* X_2. Then, we get the model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Fit the model, and test against $\beta_2 = 0$, if it is significant (by t-value), then retain X_1, X_2 in the model and compute the residual $\mathbf{e}$; if it is insignificant, stop the whole procedure and return the model $Y = \beta_0 + \beta_1 X_1$.

(c) **step 3**: From the remaining variables $X_3, \cdots, X_q$, choose the one with the **highest** correlation with $\mathbf{e}$, *say it is* X_3. Then, we get the model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$. Fit the model, and test against $\beta_3 = 0$, if it is significant (by t-value), then retain X_1, X_2, X_3 in the model and compute the residual $\mathbf{e}$; if it is insignificant, stop the whole procedure and return the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.

(d) $\vdots$

(e) $\vdots$

(f) **step q**: From the remaining variables X_q, choose the one with the **highest** correlation with $\mathbf{e}$, *say it is* X_q. Then, we get the model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q$. Fit the model, and test against $\beta_q = 0$, if it is significant (by t-value), then retain $X_1, \cdots, X_q$ in the model and compute the residual $\mathbf{e}$; if it is insignificant, stop the whole procedure and return the model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{q-1} X_{q-1}$.

If all the $q+1$ models are significant, then how do we choose one model in the end?

8.5 Variable Selection Procedures

Backward Elimination Procedure

The backward elimination procedure starts with the **full equation** and successively drops one variable at a time. The variables are dropped on the basis of their contribution to the **reduction of error sum of squares**. The first variable deleted is the one with the smallest contribution to the reduction of error sum of squares. This is equivalent to deleting the variable which has the smallest t-Test in the equation. If all the t-Tests are **significant**, the full set of variables is retained in the equation.

Assuming that there are one or more variables that have insignificant t-Tests, the procedure operates by dropping the variable with the smallest **insignificant** t-Test. The equation with the remaining $q-1$ variables is then fitted and the t-Tests for the new regression coefficients are examined. The procedure is terminated when all the t-Tests are significant or all variables have been deleted. In most backward elimination algorithms the cutoff value for the t-Test is set high so that the procedure runs through the whole set of variables, that is, starting with the q-variable equation and ending up with an equation containing only the constant term. The backward elimination procedure involves fitting at most $q+1$ regression equations

Then, among the q+1 equations, which one should we choose?

8.5 Variable Selection Procedures

Backward Elimination Procedure

(a) **Step 1**: Start from the full model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q$. Fit the model and obtain the t-statistics and identify the variable with the smallest t-value which is insignificant, *say it is* X_1. Then, we drop X_1 from the model; However, if all the t-values are significant, we stop the whole procedure and return the model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q$.

(b) **step 2**: Then consider the model: $Y = \beta_0 + \beta_2 X_2 + \cdots + \beta_q X_q$. Fit the model and obtain the t-statistics and identify the variable with the smallest t-value which is insignificant, *say it is* X_2. Then, we drop X_2 from the model; However, if all the t-values are significant, we stop the whole procedure and return the model $Y = \beta_0 + \beta_2 X_2 + \cdots + \beta_q X_q$.

(c) **step 3**: Then consider the model: $Y = \beta_0 + \beta_3 X_3 + \cdots + \beta_q X_q$. Fit the model and obtain the t-statistics and identify the variable with the smallest t-value which is insignificant, *say it is* X_3. Then, we drop X_3 from the model; However, if all the t-values are significant, we stop the whole procedure and return the model $Y = \beta_0 + \beta_3 X_3 + \cdots + \beta_q X_q$.

(d) $\vdots$

(e) $\vdots$

(f) **step q**: Finally, consider the model: $Y = \beta_0 + \beta_q X_q$. Fit the model and obtain the t-statistics and identify the variable with the smallest t-value which is insignificant, *say it is* X_q. Then, we drop X_q from the model; However, if all the t-values are significant, we stop the whole procedure and return the model $Y = \beta_0 + \beta_q X_q$.

If all the $q+1$ models are significant, then how do we choose one model in the end?

8.5 Variable Selection Procedures

Stepwise Method (optional)

The stepwise method is essentially a forward selection procedure but with the added proviso that at each stage the possibility of deleting a variable, as in backward elimination, is considered. In this procedure a variable that entered in the earlier stages of selection may be eliminated at later stages. The calculations made for inclusion and deletion of variables are the same as FS and BE procedures. Often, different levels of significance are assumed for inclusion and exclusion of variables from the equation.

AIC and BIC both can be used for setting up stepwise procedures (forward selection and backward elimination). For forward selection one starts with a constant as the fitting term, and adds variables to the model. The procedure is terminated, when addition of a variable causes no reduction of AIC (BIC). In the backward procedure, we start with the full model (containing all the variables) and drop variables successively. The procedure is terminated when dropping a variable does not lead to any further reduction in the criteria.

The stepwise procedure based on information criteria differs in a major way from the procedures based on the t-statistic that gauges the significance of a variable. The information-based procedures are driven by all the variables in the model. The termination of the procedure is based solely on the decrease of the criterion, and not on the statistical significance of the entering or departing variable.

8.5 Variable Selection Procedures

Remarks on these procedures

The variable selection procedures discussed above should be used with caution. These procedures should not be used mechanically to determine the "best" variables. The order in which the variables enter or leave the equation in variable selection procedures should not be interpreted as reflecting the relative importance of the variables. If these caveats are kept in mind, the variable selection procedures are useful tools for variable selection in **noncollinear** situations. **All three procedures will give nearly the same selection of variables with noncollinear data**. They entail much less computing than that in the analysis of all possible equations.

Several stopping rules have been proposed for the variable selection procedures. A stopping rule that has been reported to be quite effective is as follows:

- In FS: Stop if minimum t-Test is less than 1.
- In BE: Stop if minimum t-Test is greater than 1.

For example, at one step, we fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, then the **minimum t-Test** refers to the smallest value among t_1, t_2, t_3, the t-statistic for estimated $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$, respectively.

8.5 Variable Selection Procedures

Remarks on these procedures

We recommend the BE procedure over FS procedure for variable selection. One obvious reason is that in the BE procedure the equation with the full variable set is calculated and available for inspection even though it may not be used as the final equation. Although **we do not recommend the use of variable selection procedures in a collinear situation**, the BE procedure is better able to handle collinearity than the FS procedure.

In an application of variable selection procedures several equations are generated, each equation containing a different number of variables. The various equations generated can then be evaluated using a statistic such as C_p, RMS, AIC, or BIC. The residuals for the various equations should also be examined. Equations with unsatisfactory residual plots are rejected. Only a total and comprehensive analysis will provide an adequate selection of variables and a useful regression equation. This approach to variable selection is illustrated by the following example in Section 8.6.

8.6. A study of Supervisor Performance

8.6 A study of Supervisor Performance

To illustrate variable selection procedures in a noncollinear situation, consider the Supervisor Performance data discussed in Chapter 3. A regression equation was needed to study the qualities that led to the characterization of good supervisors by the people being supervised. The equation is to be constructed in an attempt to understand the supervising process and the relative importance of the different variables. In terms of the use for the regression equation, this would imply that we want accurate estimates of the regression coefficients, in contrast to an equation that is to be used only for prediction. The variables in the problem are given in Table 3.2.

The VIFs resulting from regressing Y on $X_1, X_2, \cdots, X_6$ are

$$\text{VIF}_1 = 2.7, \quad \text{VIF}_2 = 1.6, \quad \text{VIF}_3 = 2.3,$$

$$\text{VIF}_4 = 3.1, \quad \text{VIF}_5 = 1.2, \quad \text{VIF}_6 = 2.0.$$

The range of the VIFs (1.2 to 3.1) shows that collinearity is not a problem for these data. The same picture emerges if we examine the eigenvalues of the correlation matrix of the data (Table 8.1). The eigenvalues of the correlation matrix are

$$\lambda_1 = 3.169, \quad \lambda_2 = 1.006, \quad \lambda_3 = 0.763,$$

$$\lambda_4 = 0.553, \quad \lambda_5 = 0.317, \quad \lambda_6 = 0.192.$$

The sum of the reciprocals of the eigenvalues is 12.8. Since none of the eigenvalues are small (the condition number is 4.1) and the sum of the reciprocals of the eigenvalues is only about twice the number of variables, we conclude that the data in the present example are not seriously collinear and we can apply the variable selection procedures just described.

Table 8.1

8.6 A study of Supervisor Performance

Table 8.1 Correlation Matrix for the Supervisor Performance Data in Table 3.3

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1.000					
X_2	0.558	1.000				
X_3	0.597	0.493	1.000			
X_4	0.669	0.445	0.640	1.000		
X_5	0.188	0.147	0.116	0.377	1.000	
X_6	0.225	0.343	0.532	0.574	0.283	1.000

The result of forward selection procedure is given in Table 8.2. For successive equations we show the variables present, the RMS, and the value of the C_p statistic.

Forward Selection

Table 8.2 Variables Selected by the Forward Selection Method

Variables in Equation	min($\|t\|$)	RMS	C_p	p	Rank	AIC	BIC
X_1	7.74	6.993	1.41	2	1	118.63	121.43
X_1X_3	1.57	6.817	1.11	3	1	118.00	122.21
$X_1X_3X_6$	1.29	6.734	1.60	4	1	118.14	123.74
$X_1X_3X_6X_2$	0.59	6.820	3.28	5	1	119.73	126.73
$X_1X_3X_6X_2X_4$	0.47	6.928	5.07	6	1	121.45	129.86
$X_1X_3X_6X_2X_4X_5$	0.26	7.068	7.00	7	–	123.36	133.17

8.6 A study of Supervisor Performance

Forward Selection

For instance, among all the models with 2 predictors, the model $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$ has the smallest RMS

The column labeled Rank shows the rank of the subset obtained by FS relative to best subset (on the basis of RMS) of the same size. The value of p is the number of predictor variables in the equation, including a constant term. Two stopping rules are used:

1. **Stop if minimum absolute t-Test is less than $t_{0.05}(n-p)$.**
2. **Stop if minimum absolute t-Test is less than 1.**

The first rule is more stringent and terminates with variables X_1 and X_3. The second rule is less stringent and terminates with variables X_1, X_3, and X_6.

Backward Elimination

The results of applying the BE procedure are presented in Table 8.3. They are identical in structure to Table 8.2. For the BE we will use the stopping rules:

1. **Stop if minimum absolute t-Test is greater than $t_{0.05}(n-p)$.**
2. **Stop if minimum absolute t-Test is greater than 1.**

With the first stopping rule the variables selected are X_1 and X_3. With the second stopping rule the variables selected are X_1, X_3, and X_6.

Table 8.3

8.6 A study of Supervisor Performance

Table 8.3 Variables Selected by Backward Elimination Method

Variables in Equation	min($\|t\|$)	RMS	C_p	p	Rank	AIC	BIC
$X_1X_2X_3X_4X_5X_6$	0.26	7.068	7.00	7	–	123.36	133.17
$X_1X_2X_3X_4X_6$	0.47	6.928	5.07	6	1	121.45	129.86
$X_1X_2X_3X_6$	0.59	6.820	3.28	5	1	119.73	126.73
$X_1X_3X_6$	1.29	6.734	1.60	4	1	118.14	123.74
X_1X_3	1.57	6.817	1.11	3	1	118.00	122.21
X_1	7.74	6.993	1.41	2	1	118.63	121.43

The FS and BE give identical equations for this problem, but this is not always the case. To describe the supervisor performance, the equation

$$Y = 13.58 + 0.62X_1 + 0.31X_3 - 0.19X_6$$

is chosen. The residual plots (not shown) for this equation are satisfactory. Since the present problem has only six variables, the total number of equations that can be fitted which contain at least one variable is 63. The C_p values for all 63 equations are shown in Table 8.4. The C_p values are plotted against p in Figure 8.1. The best subsets of variables based on C_p values are given in Table 8.5.

Table 8.4, 8.5
Figure 8.1

8.6 A study of Supervisor Performance

Table 8.4 Values of C_p Statistic (All Possible Equations)

Variables	C_p	Variables	C_p	Variables	C_p	Variables	C_p
1	1.41	1 5	3.41	1 6	3.33	1 5 6	5.32
2	44.40	2 5	45.62	2 6	46.39	2 5 6	47.91
1 2	3.26	1 2 5	5.26	1 2 6	5.22	1 2 5 6	7.22
3	26.56	3 5	27.94	3 6	24.82	3 5 6	25.02
1 3	1.11	1 3 5	3.11	1 3 6	1.60	1 3 5 6	3.46
2 3	26.96	2 3 5	28.53	2 3 6	24.62	2 3 5 6	25.11
1 2 3	2.51	1 2 3 5	4.51	1 2 3 6	3.28	1 2 3 5 6	5.14
4	30.06	4 5	31.62	4 6	27.73	4 5	29.50
1 4	3.19	1 4 5	5.16	1 4 6	4.70	1 4 5 6	6.69
2 4	29.20	2 4 5	30.82	2 4 6	25.91	2 4 5 6	27.74
1 2 4	4.99	1 2 4 5	6.97	1 2 4 6	6.63	1 2 4 5 6	8.61
3 4	23.25	3 4 5	25.23	3 4 6	16.50	3 4 5 6	18.42
1 3 4	3.09	1 3 4 5	5.09	1 3 4 6	3.35	1 3 4 5 6	5.29
2 3 4	24.56	2 3 4 5	26.53	2 3 4 6	17.57	2 3 4 5 6	19.51
1 2 3 4	4.49	1 2 3 4 5	6.48	1 2 3 4 6	5.07	1 2 3 4 5 6	7
5	57.91	6	57.95	5 6	58.76		

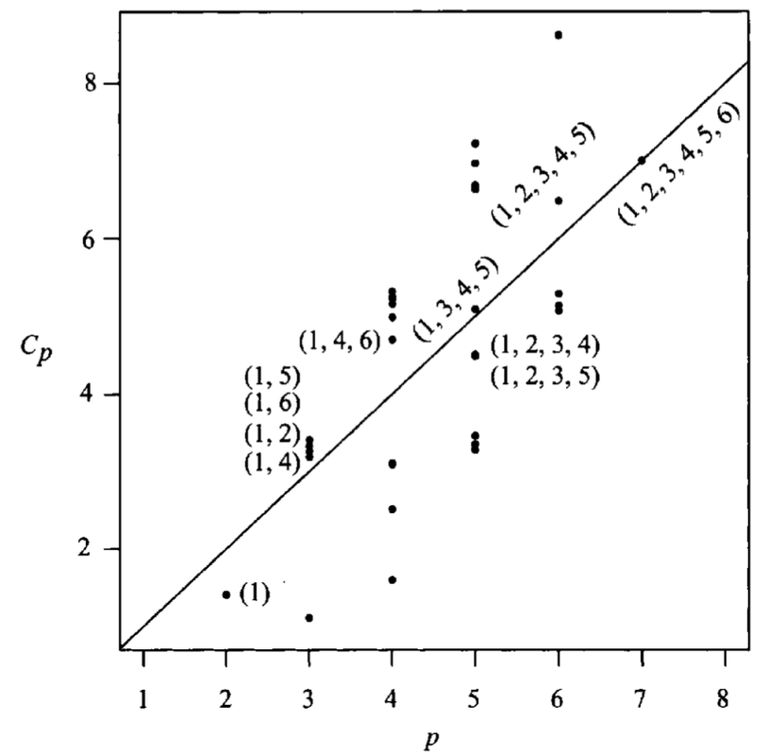

Figure 8.1 Supervisor's Performance data: Scatter plot of C_p versus p for subsets with $C_p < 10$.

Table 8.5 Variables Selected on the Basis of C_p Statistic

Variables in Equation	min($\|t\|$)	RMS	C_p	p	Rank	AIC	BIC
X_1	7.74	6.993	1.41	2	1	118.63	121.43
X_1X_4	0.47	7.093	3.19	3	2	120.38	124.59
$X_1X_4X_6$	0.69	7.163	4.70	4	5	121.84	127.45
$X_1X_3X_4X_5$	0.07	7.080	5.09	5	6	121.97	127.97
$X_1X_2X_3X_4X_5$	0.11	7.139	6.48	6	4	123.24	131.65
$X_1X_2X_3X_4X_5X_6$	0.26	7.068	7.00	7	–	133.17	133.17

8.6 A study of Supervisor Performance

It is seen that the subsets selected by C_p are different from those arrived at by the variable selection procedures as well as those selected on the basis of residual mean square. This anomaly suggests an important point concerning the C_p statistic that we should bear in mind. For applications of the C_p statistic, an estimate of σ^2 is required. Usually, the estimate of σ^2 is obtained from the residual sum of squares from the full model. If the full model has a large number of variables with **no explanatory power** (i.e., population regression coefficients are zero), the estimate of σ^2 from the residual sum of squares for the full model would be large. The loss in degrees of freedom for the divisor would not be balanced by a reduction in the error sum of squares. If $\hat{\sigma}^2$ is large, then the value of C_p is small. For C_p to work properly, a good estimate of σ^2 must be available. **When a good estimate of σ^2 is not available, C_p is of only limited usefulness.**

In our present example, the RMS for the full model with six variables is larger than the RMS for the model with three variables X_1, X_3, X_6. Consequently, the C_p values are distorted and not very useful in variable selection in the present case. The type of situation we have described can be spotted by looking at the RMS for different values of p. RMS will at first tend to decrease with p, but increase at later stages. This behavior indicates that the latter variables are not contributing significantly to the reduction of error sum of squares. Useful application of C_p requires a parallel monitoring of RMS to avoid distortions.

8.6 A study of Supervisor Performance

Values of AIC and BIC for forward selection and backward elimination is given in Tables 8.2 and 8.3. The lowest value of AIC (118.00) is obtained for X_1 and X_3. If we regard models with AIC within 2 to be equivalent, then X_1, X_1X_3, $X_1X_3X_6$, and $X_1X_3X_6X_2$ should be considered. Among these four candidate models we can pick one of them. The lowest value of BIC (121.43) is attained by X_1. There is only one other model (X_1X_3) whose BIC lies within 2 units. It should be noted that BIC selects models with smaller number of variables because of its penalty function. Variable selection should not be done mechanically. In many situations there may not be a "best model" or a "best set of variables." The aim of the analysis should be to identify all models of high equal adequacy.