CHAPTER 11

VARIABLE SELECTION PROCEDURES

11.1 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.

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?

Regression Analysis by Example, Fourth Edition. By Samprit Chatterjee and Ali S. Hadi

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.

11.2 FORMULATION OF THE PROBLEM

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

$$y_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij} + \varepsilon_i, \tag{11.1}$$

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 $X_1, X_2, \ldots, X_p$ and those deleted by $X_{p+1}, X_{p+2}, \ldots, X_q$. 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 $(\beta_0, \beta_1, \ldots, \beta_q)$ nonzero.
2. The model has $\beta_0, \beta_1, \ldots, \beta_p$ nonzero, but $\beta_{p+1}, \beta_{p+2}, \ldots, \beta_q$ zero.

Suppose that instead of fitting (11.1) we fit the subset model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i. \tag{11.2}$$

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.

11.3 CONSEQUENCES OF VARIABLES DELETION

Denote the estimates of the regression parameters by $\hat{\beta}_0^*, \hat{\beta}_1^*, \ldots, \hat{\beta}_q^*$ when the model (11.1) is fitted to the full set of variables $X_1, X_2, \ldots, X_q$. Denote the

estimates of the regression parameters by $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ when the model (11.2) 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 $(x_{i1}, x_{i2}, \ldots, x_{iq})$. The results can now be summarized as follows (a summary using matrix notation is given in the Appendix to this chapter): $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are biased estimates of $\beta_0, \beta_1, \ldots, \beta_p$ unless the remaining β's in the model $(\beta_{p+1}, \beta_{p+2}, \ldots, \beta_q)$ are zero or the variables $X_1, X_2, \ldots, X_p$ are orthogonal to the variable set $(X_{p+1}, X_{p+2}, \ldots, X_q)$. The estimates $\hat{\beta}_0^*, \hat{\beta}_1^*, \ldots, \hat{\beta}_p^*$ have less precision than $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$; that is,

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

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. 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.

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,

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

The conditions for MSE($\hat{y}_i$) to be smaller than Var($\hat{y}_i^*$) are identical to the conditions for MSE($\hat{\beta}_j$) to be smaller than Var($\hat{\beta}_j^*$), which we have already stated. For further details, refer to Chatterjee and Hadi (1988).

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.

The reader is referred to Sections 3.5, 4.12, and 4.13 for further elaboration on the interpretation of regression coefficients and the role of variables in regression modeling.

11.4 USES OF REGRESSION EQUATIONS

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

11.4.1 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.

11.4.2 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.

11.4.3 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 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.

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.

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.

11.5 CRITERIA FOR EVALUATING EQUATIONS

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. An exhaustive list of criteria is found in Hocking (1976).

11.5.1 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

$$\text{RMS}_p = \frac{\text{SSE}_p}{n-p} . \tag{11.3}$$

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.

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

$$R_p^2 = 1 - (n-p)\frac{\text{RMS}_p}{(\text{SST})} \tag{11.4}$$

and

$$R_{ap}^2 = 1 - (n-1)\frac{\text{RMS}_p}{(\text{SST})}\,, \tag{11.5}$$

where

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

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

11.5.2 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

$$J_p = \frac{1}{\sigma^2}\sum_{i=1}^{n} \text{MSE}(\hat{y}_i), \tag{11.6}$$

where $\text{MSE}(\hat{y}_i)$ 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 $\text{MSE}(\hat{y}_i)$ has two components, the variance of prediction arising from estimation, and a bias component arising from the deletion of variables.

To estimate J_p, Mallows (1973) uses the statistic

$$C_p = \frac{\text{SSE}_p}{\hat{\sigma}^2} + (2p - n), \tag{11.7}$$

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 11.10. A very thorough treatment of the C_p statistic is given in Daniel and Wood (1980).

11.5.3 Information Criteria: Akaike and Other Modified Forms

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 Criteria (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 Section 3.9.2. AIC for a p-term equation (a constant, and $(p-1)$ variables) is given by

$$\text{AIC}_p = n \ln(\text{SSE}_p/n) + 2p. \tag{11.8}$$

The models with smaller AIC are preferred.

We can see from (11.8) that for two models with similar SSE, AIC penalizes the model that has 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.

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 Section 3.9). 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 a complete data cases (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.

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

$$\text{BIC}_p = n \ln(\text{SSE}_p/n) + p(\ln n). \tag{11.9}$$

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.

Another modification of AIC to avoid overfitting is the bias corrected version, AIC^c, proposed by Hurvich and Tsai (1989), is given by

$$\text{AIC}_p^c = \text{AIC}_p + \frac{2(p+2)(p+3)}{n-p-3}. \tag{11.10}$$

The correction to AIC in (11.10) is small for large n and moderate p. The correction is large when n is small and p large. One should never fit a large and complex

model with a small number of observations. In general the correction to AIC will be minor, and we will not discuss AIC^c further. To guard against overfitting in our analysis we will examine BIC.

11.6 MULTICOLLINEARITY AND VARIABLE SELECTION

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 multicollinearity.

2. The predictor variables are collinear; that is, the data are highly multicollinear.

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 9, the presence of small eigenvalues indicates collinearity. If the condition number[1] 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.

11.7 EVALUATING ALL POSSIBLE EQUATIONS

The first procedure described is very direct and applies equally well to both collinear and noncollinear 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 may be 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

[1]Recall from Chapter 9 that the condition number is defined by $\kappa = \sqrt{\lambda_{max}/\lambda_{min}}$ where λ_{max} and λ_{min} are the maximum and minimum eigenvalues of the matrix of correlation coefficients.

fitting 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, autocorrelation, 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.

When the number of variables is large, the evaluation of all possible equations may not be practically feasible. Certain shortcuts have been suggested (Furnival and Wilson, 1974; La Motte and Hocking, 1970) 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 11.8. These procedures are quite efficient with noncollinear data. We do not, however, recommend them for collinear data.

11.8 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* procedure (FS), and (2) the *backward elimination* procedure (BE). There is also a very popular modification of the FS procedure called the *stepwise* method. The three procedures are described and compared below.

11.8.1 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.

11.8.2 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

11.8.3 Stepwise Method

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.

Most of the currently available software do not automatically produce AIC or BIC. They all, however, provide SSE, from which it is easy to compute (11.8) and (11.9) the information criteria.

11.9 GENERAL REMARKS ON VARIABLE SELECTION METHODS

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.

In the following example we illustrate the effect of different stopping rules in variable selection.

We recommend the BE procedure over FS procedure for variable selection. One obvious reason is that in 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 multicollinearity than the FS procedure (Mantel, 1970).

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.

Table 11.1 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

11.10 A STUDY OF SUPERVISOR PERFORMANCE

To illustrate variable selection procedures in a noncollinear situation, consider the Supervisor Performance data discussed in Section 3.3. 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 data are shown in Table 3.3 and can also be obtained from the book's Web site.[2]

The VIFs resulting from regressing Y on $X_1, X_2, \ldots, X_6$ 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 11.1). 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.

The result of forward selection procedure is given in Table 11.2. For successive equations we show the variables present, the RMS, and the value of the C_p statistic.

[2]http://www.ilr.cornell.edu/~hadi/RABE4

Table 11.2 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

The column labeled Rank shows the rank of the subset obtained by FS relative to best subset (on the basis of RMS) of 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.

The results of applying the BE procedure are presented in Table 11.3. They are identical in structure to Table 11.2. 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. The FS and BE give identical equations for this problem, but this is not always the case (an example is given in Section 11.12). 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 11.4. The C_p values are plotted against p in Figure 11.1. The best subsets of variables based on C_p values are given in Table 11.5.

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 the reader 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

Table 11.3 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

Table 11.4 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		

Table 11.5 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

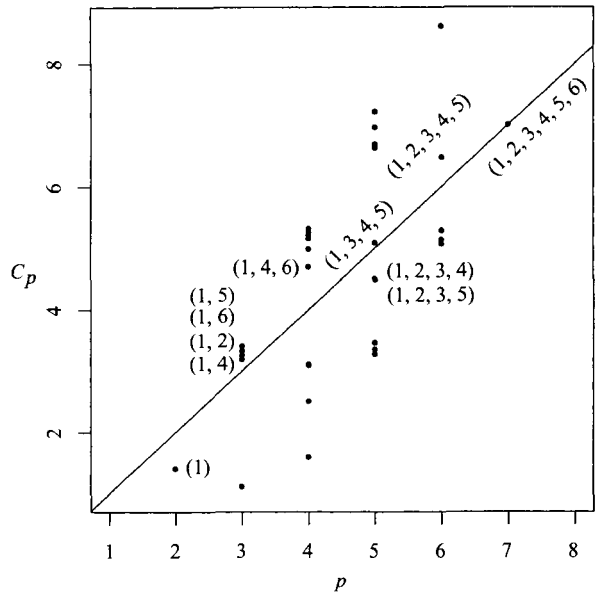

Figure 11.1 Supervisor's Performance Data: Scatter plot of C_p versus p for subsets with $C_p < 10$.

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.

Values of AIC and BIC for forward selection and backward elimination is given in Tables 11.2 and 11.3. 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, $X_1X_3X_6$, $X_1X_3X_6X_2$, 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.

11.11 VARIABLE SELECTION WITH COLLINEAR DATA

In Chapter 9 it was pointed out that serious distortions are introduced in standard analysis with collinear data. Consequently, we recommend a different set of procedures for selecting variables in these situations. Collinearity is indicated when the correlation matrix has one or more small eigenvalues. With a small number of collinear variables we can evaluate all possible equations and select an equation by methods that have already been described. But with a larger number of variables this method is not feasible.

Two different approaches to the problem have been proposed. The first approach tries to break down the collinearity of the data by deleting variables. The collinear structure present in the variables is revealed by the eigenvectors corresponding to the very small eigenvalues (see Chapters 9 and 10). Once the collinearities are identified, a set of variables can then be deleted to produce a reduced noncollinear data set. We can then apply the methods described earlier. The second approach uses ridge regression as the main tool. We assume that the reader is familiar with the basic terms and concepts of ridge regression (Chapter 10). The first approach (by judicious dropping of correlated variables) is the one that is almost always used in practice.

11.12 THE HOMICIDE DATA

In a study investigating the role of firearms in accounting for the rising homicide rate in Detroit, data were collected for the years 1961–1973. The data are reported in Gunst and Mason (1980), p. 360. The response variable (the homicide rate) and the predictor variables believed to influence or be related to the rise in the homicide rate are defined in Table 11.6 and given in Tables 11.7 and 11.8. The data can also be found in the book's Web site.

We use these data to illustrate the danger of mechanical variable selection procedures, such as the FS and BE, in collinear situations. We are interested in fitting the model

$$H = \beta_0 + \beta_1 G + \beta_2 M + \beta_3 W + \varepsilon.$$

In terms of the centered and scaled version of the variables, the model becomes

$$\tilde{H} = \theta_1 \tilde{G} + \theta_2 \tilde{M} + \theta_3 \tilde{W} + \varepsilon'. \tag{11.11}$$

The OLS results are shown in Table 11.9. Can the number of predictor variables in this model be reduced? If the standard assumptions hold, the small t-test for the variable G (0.68) would indicate that the corresponding regression coefficient

Table 11.6 Homicide Data: Description of Variables

Variable	Symbol	Description
1	FTP	Number of full-time police per 100,000 population
2	UEMP	Percent of the population unemployed
3	M	Number of manufacturing workers (in thousands)
4	LIC	Number of handgun licenses issued per 100,000 population
5	GR	No. of handgun registration issued per 100,000 population
6	CLEAR	Percent of homicides cleared by arrest
7	W	Number of white males in the population
8	NMAN	Number of nonmanufacturing workers (in thousands)
9	G	Number of government workers (in thousands)
10	HE	Average hourly earnings
11	WE	Average weekly earnings
12	H	Number of homicides per 100,000 population

Table 11.7 First Part of the Homicides Data

Year	FTP	UNEMP	M	LIC	GR	CLEAR
1961	260.35	11.0	455.5	178.15	215.98	93.4
1962	269.80	7.0	480.2	156.41	180.48	88.5
1963	272.04	5.2	506.1	198.02	209.57	94.4
1964	272.96	4.3	535.8	222.10	231.67	92.0
1965	272.51	3.5	576.0	301.92	297.65	91.0
1966	261.34	3.2	601.7	391.22	367.62	87.4
1967	268.89	4.1	577.3	665.56	616.54	88.3
1968	295.99	3.9	596.9	1131.21	1029.75	86.1
1969	319.87	3.6	613.5	837.80	786.23	79.0
1970	341.43	7.1	569.3	794.90	713.77	73.9
1971	356.59	8.4	548.8	817.74	750.43	63.4
1972	376.69	7.7	563.4	583.17	1027.38	62.5
1973	390.19	6.3	609.3	709.59	666.50	58.9

Source: Gunst and Mason (1980), p. 360

Table 11.8 Second Part of the Homicide Data

Year	W	NMAN	G	HE	WE	H
1961	558724	538.1	133.9	2.98	117.18	8.60
1962	538584	547.6	137.6	3.09	134.02	8.90
1963	519171	562.8	143.6	3.23	141.68	8.52
1964	500457	591.0	150.3	3.33	147.98	8.89
1965	482418	626.1	164.3	3.46	159.85	13.07
1966	465029	659.8	179.5	3.60	157.19	14.57
1967	448267	686.2	187.5	3.73	155.29	21.36
1968	432109	699.6	195.4	2.91	131.75	28.03
1969	416533	729.9	210.3	4.25	178.74	31.49
1970	401518	757.8	223.8	4.47	178.30	37.39
1971	398046	755.3	227.7	5.04	209.54	46.26
1972	373095	787.0	230.9	5.47	240.05	47.24
1973	359647	819.8	230.2	5.76	258.05	52.33

Source: Gunst and Mason (1980), p. 360

Table 11.9 Homicide Data: The OLS Results From Fitting Model (11.11)

Variable	Coefficient	s.e.	t-test	VIF
G	0.235	0.345	0.68	42
M	−0.405	0.090	−4.47	3
W	−1.025	0.378	−2.71	51
$n = 13$	$R^2 = 0.975$	$R_a^2 = 0.966$	$\hat{\sigma} = 0.0531$	$d.f. = 9$

is insignificant and G can be omitted from the model. Let us now apply the forward selection and the backward elimination procedures to see which variables are selected. The regression output that we need to implement the two methods on the standardized versions of the variables are summarized in Table 11.10. In this table we give the estimated coefficients, their t-tests, and the adjusted squared multiple correlation coefficient, R_a^2 for each model for comparison purposes.

The first variable to be selected by the FS is G because it has the largest t-test among the three models that contain a single variable (Models (a) to (c) in Table 11.10). Between the two candidates for the two-variable models (Models (d) and (e)), Model (d) is better than Model (e). Therefore, the second variable to enter the equation is M. The third variable to enter the equation is W (Model (f)) because it has a significant t-test. Note, however, the dramatic change of the significance of G in Models (a), (d), and (f). It was highly significant coefficient in Models (a) and (d), but became insignificant in Model (f). Collinearity is a suspect!

Table 11.10 Homicide Data: The Estimated Coefficients, Their t-tests, and the Adjusted Squared Multiple Correlation Coefficient, R_a^2

	Model						
Variable	(a)	(b)	(c)	(d)	(e)	(f)	(g)
G: Coeff.	0.96			1.15	0.87	0.24	
t-test	11.10			11.90	1.62	0.68	
M: Coeff.		0.55		−0.27		−0.40	−0.43
t-test		2.16		−2.79		−4.47	−5.35
W: Coeff.			−0.95		−0.09	−1.02	−1.28
t-test			−9.77		−0.17	−2.71	−15.90
R_a^2	0.91	0.24	0.89	0.95	0.90	0.97	0.97

The BE methods starts with the three-variable Model (f). The first variable to leave is G (because it has the lowest t-test), which leads to Model (g). Both M and W in Model (g) have significant t-tests and the BE procedure terminates.

Observe that the first variable eliminated by the BE (G) is the same as the first variable selected by the FS. That is, the variable G, which was selected by the FS as the most important of the three variables, was regarded by the BE as the least important! Among other things, the reason for this anomalous result is collinearity. The eigenvalues of the correlation matrix, $\lambda_1 = 2.65$, $\lambda_2 = 0.343$, and $\lambda_3 = 0.011$, give a large condition number ($\kappa = 15.6$). Two of the three variables (G and W) have large VIF (42 and 51). The sum of the reciprocals of the eigenvalues is also very large (96). In addition to collinearity, since the observations were taken over time (for the years 1961–1973), we are dealing with time series data here. Consequently, the error terms can be autocorrelated (see Chapter 8). Examining the pairwise scatter plots of the data will reveal other problems with the data.

This example shows clearly that automatic applications of variable selection procedure in multicollinear data can lead to the selection of a wrong model. In Sections 11.13 and 11.14 we make use of ridge regression for the process of variable selection in multicollinear situations.

11.13 VARIABLE SELECTION USING RIDGE REGRESSION

One of the goals of ridge regression is to produce a regression equation with stable coefficients. The coefficients are stable in the sense that they are not affected by slight variations in the estimation data. The objectives of a good variable selection procedure are (1) to select a set of variables that provides a clear understanding of the process under study, and (2) to formulate an equation that provides accurate

forecasts of the response variable corresponding to values of the predictor variables not included in the study. It is seen that the objectives of a good variable selection procedure and ridge regression are very similar and, consequently, one (ridge regression) can be employed to accomplish the other (variable selection).

The variable selection is done by examining the ridge trace, a plot of the ridge regression coefficients against the ridge parameter k. For a collinear system, the characteristic pattern of ridge trace has been described in Chapter 10. The ridge trace is used to eliminate variables from the equation. The guidelines for elimination are:

1. Eliminate variables whose coefficients are stable but small. Since ridge regression is applied to standardized data, the magnitude of the various coefficients are directly comparable.

2. Eliminate variables with unstable coefficients that do not hold their predicting power, that is, unstable coefficients that tend to zero.

3. Eliminate one or more variables with unstable coefficients. The variables remaining from the original set, say p in number, are used to form the regression equation.

At the end of each of the above steps, we refit the model that includes the remaining variables before we proceed to the next step.

The subset of variables remaining after elimination should be examined to see if collinearity is no longer present in the subset. We illustrate this procedure by an example.

11.14 SELECTION OF VARIABLES IN AN AIR POLLUTION STUDY

McDonald and Schwing (1973) present a study that relates total mortality to climate, socioeconomic, and pollution variables. Fifteen predictor variables selected for the study are listed in Table 11.11. The response variable is the total age-adjusted mortality from all causes. We will not comment on the epidemiological aspects of the study, but merely use the data as an illustrative example for variable selection. A very detailed discussion of the problem is presented by McDonald and Schwing in their paper and we refer the interested reader to it for more information.

The original data are not available to us, but the correlation matrix of the response and the 15 predictor variables is given in Table 11.12. It is not a good practice to perform the analysis based only on the correlation matrix because without the original data we will not be able to perform diagnostics checking which is necessary in any thorough data analysis. To start the analysis we shall assume that the standard assumptions of the linear regression model hold. As can be expected from the nature of the variables, some of them are highly correlated with each other. The evidence of collinearity is clearly seen if we examine the eigenvalues of the correlation

Table 11.11 Description of Variables, Means, and Standard Deviations, SD ($n = 60$)

Variable	Description	Mean	SD
X_1	Mean annual precipitation (inches)	37.37	9.98
X_2	Mean January temperature (degrees Fahrenheit)	33.98	10.17
X_3	Mean July temperature (degrees Fahrenheit)	74.58	4.76
X_4	Percent of population over 65 years of age	8.80	1.46
X_5	Population per household	3.26	0.14
X_6	Median school years completed	10.97	0.85
X_7	Percent of housing units that are sound	80.92	5.15
X_8	Population per square mile	3876.05	1454.10
X_9	Percent of nonwhite population	11.87	8.92
X_{10}	Percent employment in white-collar jobs	46.08	4.61
X_{11}	Percent of families with income under $3000	14.37	4.16
X_{12}	Relative pollution potential of hydrocarbons	37.85	91.98
X_{13}	Relative pollution potential of oxides of nitrogen	22.65	46.33
X_{14}	Relative pollution potential of sulfur dioxide	53.77	63.39
X_{15}	Percent relative humidity	57.67	5.37
Y	Total age-adjusted mortality from all causes.	940.36	62.21

matrix. The eigenvalues are

$$\lambda_1 = 4.5272, \quad \lambda_6 = 0.9605, \quad \lambda_{11} = 0.1665,$$

$$\lambda_2 = 2.7547, \quad \lambda_7 = 0.6124, \quad \lambda_{12} = 0.1275,$$

$$\lambda_3 = 2.0545, \quad \lambda_8 = 0.4729, \quad \lambda_{13} = 0.1142,$$

$$\lambda_4 = 1.3487, \quad \lambda_9 = 0.3708, \quad \lambda_{14} = 0.0460,$$

$$\lambda_5 = 1.2227, \quad \lambda_{10} = 0.2163, \quad \lambda_{15} = 0.0049.$$

There are two very small eigenvalues; the largest eigenvalue is nearly 1000 times larger than the smallest eigenvalue. The sum of the reciprocals of the eigenvalues is 263, which is nearly 17 times the number of variables. The data show strong evidence of collinearity.

The initial OLS results from fitting a linear model to the centered and scaled data are given in Table 11.13. Although the model has a high R^2, some of the estimated coefficients have small t-tests. In the presence of multicollinearity, a small t-test does not necessarily mean that the corresponding variable is not important. The small t-test might be due of variance inflation because of the presence of multicollinearity. As can be seen in Table 11.13, VIF_{12} and VIF_{13} are very large.

Table 11.12 Correlation Matrix for the Variables in Table 11.11.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	Y
X_1	1.0000	.0922	.5033	.1011	.2634	−.4904	−.4903	−.0035	.4132	−.2973	.5066	−.5318	−.4873	−.1069	−.0773	.5095
X_2		1.0000	.3463	−.3981	−.2092	.1163	.0139	−.1001	.4538	.2380	.5653	.3508	.3210	−.1078	.0679	−.0300
X_3			1.0000	−.4340	.2623	−.2385	−.4155	−.0610	.5753	−.0214	.6193	−.3565	−.3377	−.0993	−.4528	.2770
X_4				1.0000	−.5091	−.1389	.0649	.1620	−.6378	−.1177	−.3098	−.0205	−.0021	.0172	.1124	−.1746
X_5					1.0000	−.3951	−.4095	−.1843	.4194	−.4257	.2599	−.3882	−.3584	−.0041	−.1357	.3573
X_6						1.0000	.5515	−.2439	−.2088	.7032	−.4033	.2868	.2244	−.2343	.1765	−.5110
X_7							1.0000	.1806	−.4091	.3376	−.6806	.3859	.3476	.1180	.1224	−.4248
X_8								1.0000	−.0057	−.0318	−.1629	.1203	.1653	.4321	−.1250	.2655
X_9									1.0000	−.0044	.7049	−.0259	.0184	.1593	−.1180	.6437
X_{10}										1.0000	−.1852	.2037	.1600	−.0685	.0607	−.2848
X_{11}											1.0000	−.1298	−.1025	−.0965	−.1522	.4105
X_{12}												1.0000	.9838	.2823	−.0202	−.1772
X_{13}													1.0000	.4094	−.0459	−.0774
X_{14}														1.0000	−.1026	.4259
X_{15}															1.0000	−.0885
Y																1.0000

Table 11.13 OLS Regression Output for the Air Pollution Data (Fifteen Predictor Variables)

Variable	Coefficient	s.e.	t-test	VIF
X_1	0.306	0.148	2.063	4.11
X_2	−0.318	0.181	−1.755	6.13
X_3	−0.237	0.146	−1.627	3.97
X_4	−0.213	0.200	−1.064	7.46
X_5	−0.232	0.152	−1.527	4.31
X_6	−0.233	0.161	−1.448	4.85
X_7	−0.052	0.146	−0.356	3.97
X_8	0.084	0.094	0.890	1.66
X_9	0.640	0.190	3.359	6.78
X_{10}	−0.014	0.123	−0.112	2.84
X_{11}	−0.010	0.216	−0.042	8.72
X_{12}	−0.979	0.724	−1.353	97.92
X_{13}	0.983	0.747	1.316	104.22
X_{14}	0.090	0.150	0.599	4.21
X_{15}	0.009	0.101	0.093	1.91
$n = 60$	$R^2 = 0.764$	$R_a^2 = 0.648$	$\hat{\sigma} = 0.073$	$d.f. = 44$

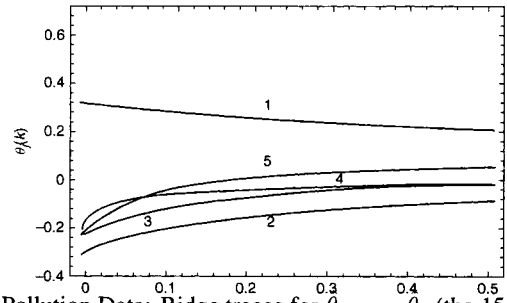

Figure 11.2 Air Pollution Data: Ridge traces for $\theta_1, \ldots, \theta_5$ (the 15-variable-model).

The ridge trace for the 15 regression coefficients are shown in Figures 11.2 to 11.4. Each Figure shows five curves. If we put all 15 curves, the graph would be quite cluttered and the curves would be difficult to trace. To make the three graphs comparable, the scale is kept the same for all graphs. From the ridge trace, we see that some of the coefficients are quite unstable and some are small regardless of the value of the ridge parameter k.

We now follow the guidelines suggested for the selection of variables in multicollinear data. Following the first criterion we eliminate variables 7, 8, 10, 11, and 15. These variables all have fairly stable coefficients, as shown by the flatness of their ridge traces, but are very small. Although variable 14 has a small coefficient

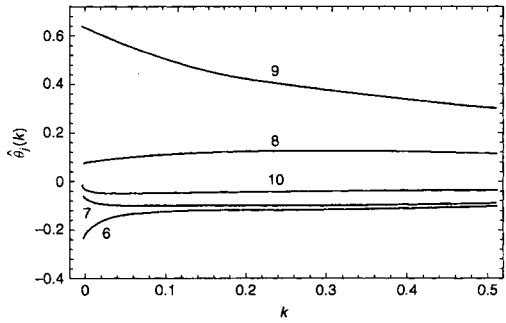

Figure 11.3 Air Pollution Data: Ridge traces for $\hat{\theta}_6, \ldots, \hat{\theta}_{10}$ (the 15-variable-model).

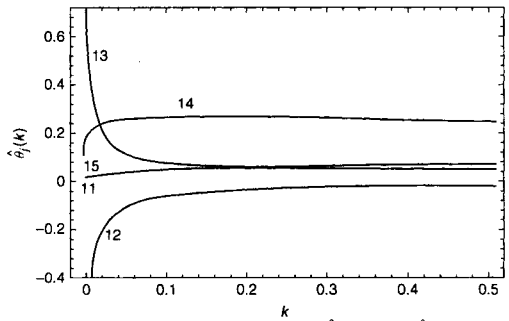

Figure 11.4 Air Pollution Data: Ridge traces for $\hat{\theta}_{11}, \ldots, \hat{\theta}_{15}$ (the 15-variable-model).

at $k = 0$ (see Table 11.13), its value increases sharply as k increases from zero. So, it should not be eliminated at this point.

We now repeat the analysis using the ten remaining variables: 1, 2, 3, 4, 5, 6, 9, 12, 13, and 14. The corresponding OLS results are given in Table 11.14. There is still an evidence of multicollinearity. The largest eigenvalue, $\lambda_1 = 3.377$, is about 600 times the smallest value $\lambda_{10} = 0.005$. The two VIFs for variable 12 and 13 are still high. The corresponding ridge traces are shown in Figures 11.5 and 11.6. Variable 14 continues to have a small coefficient at $k = 0$ but it increases as k increases from zero. So, it should be kept in the model at this stage. None of the other nine variables satisfy the first criterion.

The second criterion suggests eliminating variables with unstable coefficients that tend to zero. Examination of the ridge traces in Figures 11.5 and 11.6 shows that variables 12 and 13 fall in this category.

The OLS results for the remaining 8 variables are shown in Table 11.15. Collinearity has disappeared. Now, the largest and smallest eigenvalues are 2.886 and 0.094, which give a small condition number ($\kappa = 5.5$). The sum of the recip-

Table 11.14 OLS Regression Output for the Air Pollution Data (Ten Predictor Variables)

Variable	Coefficient	s.e.	t-test	VIF
X_1	0.306	0.135	2.260	3.75
X_2	−0.345	0.119	−2.907	2.88
X_3	−0.244	0.108	−2.256	2.39
X_4	−0.222	0.175	−1.274	6.22
X_5	−0.268	0.137	−1.959	3.81
X_6	−0.292	0.103	−2.842	2.15
X_9	0.664	0.140	4.748	3.99
X_{12}	−1.001	0.658	−1.522	88.30
X_{13}	1.001	0.673	1.488	92.40
X_{14}	0.098	0.127	0.775	3.29
$n = 60$	$R^2 = 0.760$	$R_a^2 = 0.711$	$\hat{\sigma} = 0.070$	$d.f. = 49$

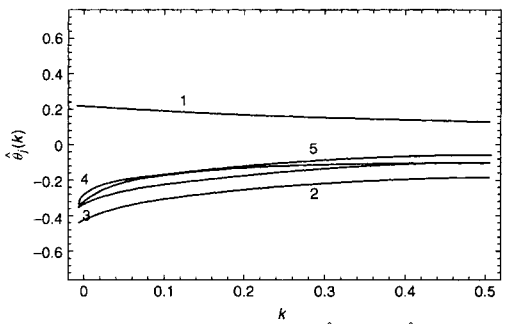

Figure 11.5 Air Pollution Data: Ridge traces for $\hat{\theta}_1, \ldots, \hat{\theta}_5$ (the ten-variable-model).

rocals of the eigenvalues is 23.5, about twice the number of variables. All values of VIF are less than 10. Since the retained variables are not collinear, we can now apply the variables selection methods for non-collinear data discussed in Sections 11.7 and 11.8. This is left as an exercise for the reader.

An alternative way of analyzing these Air Pollution data is as follows: The collinearity in the original 15 variables is actually a simple case of multicollinearity; it involves only two variables (12 and 13). So, the analysis can proceed by eliminating any one of the two variables. The reader can verify that the remaining 14 variables are not collinear. The standard variables selection procedures for non-collinear data can now be utilized. We leave this as an exercise for the reader.

In our analysis of the Air Pollution data, we did not use the third criterion, but there are situations where this criterion is needed. We should note that ridge regression was used successfully in this example as a tool for variable selection.

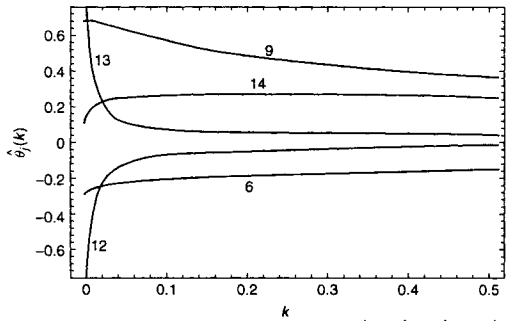

Figure 11.6 Air Pollution Data: Ridge traces for $\hat{\theta}_6$, $\hat{\theta}_9$, $\hat{\theta}_{12}$, $\hat{\theta}_{13}$ and $\hat{\theta}_{14}$ (the ten-variable-model).

Table 11.15 OLS Regression Output for the Air Pollution Data (Eight Predictor Variables)

Variable	Coefficient	s.e.	t-test	VIF
X_1	0.331	0.120	2.765	2.911
X_2	−0.351	0.106	−3.313	2.279
X_3	−0.217	0.104	−2.087	2.191
X_4	−0.155	0.163	−0.946	5.419
X_5	−0.221	0.134	−1.656	3.621
X_6	−0.270	0.102	−2.654	2.097
X_9	0.692	0.133	5.219	3.567
X_{14}	0.230	0.083	2.767	1.405
$n = 60$	$R^2 = 0.749$	$R_a^2 = 0.709$	$\hat{\sigma} = 0.070$	$d.f. = 51$

Because the variables selected at an intermediate stage were found to be non-collinear, the standard OLS was utilized.

An analysis of these data not using ridge regression has been given by Henderson and Velleman (1981). They present a thorough analysis of the data and the reader is referred to their paper for details.

Some General Comments: We hope it is clear from our discussion that variable selection is a mixture of art and science, and should be performed with care and caution. We have outlined a set of approaches and guidelines rather than prescribing a formal procedure. In conclusion, we must emphasize the point made earlier that variable selection should not be performed mechanically as an end in itself but rather as an exploration into the structure of the data analyzed, and as in all true explorations, the explorer is guided by theory, intuition, and common sense.

11.15 A POSSIBLE STRATEGY FOR FITTING REGRESSION MODELS

In the concluding section of the chapter we outline a possible sequence of steps that may be used to fit a regression model satisfactorily. Let us emphasize at the beginning that there is no single correct approach. The reader may be more comfortable with a different sequence of steps and should feel free to follow such a sequence. In almost all cases the analysis described here will lead to meaningful interpretable models useful in real-life applications.

We assume that we have a response variable Y which we want to relate to some or all of a set of variables $X_1, X_2, \ldots, X_p$. The set, $X_1, X_2, \ldots, X_p$, is often generated from external subject matter considerations. The set of variables is often large and we want to come to an acceptable reduced set. Our objective is to construct a valid and viable regression model. A possible sequence of steps are:

1. Examine the variables $(Y, X_1, X_2, \ldots, X_p)$ one at a time. This can be done by calculating the summary statistics, and also graphically by looking at histograms, dot plots, or box plots (see Chapter 4). The distributions of the values should not be too skewed, nor the range of the variables very large. Look for outliers (check for transcription errors). Make transformations to induce symmetry and reduce skewness. Logarithmic transformations are useful in this situation (see Chapter 6).

2. Construct pairwise scatter plots for each variable. When p, the number of predictor variables, is large, this may not be feasible. Pairwise scatter plots are quite informative on the relationship between two variables. A look at the correlation matrix will point out obvious collinearity problems. Delete redundant variables. Calculate the condition number of the correlation matrix to get an idea of the severity of the collinearity (Chapters 9 and 10).

3. Fit the full linear regression model. Delete variables with no significant explanatory power (insignificant t-tests). For the reduced model, examine the residuals:

   (a) Check linearity. If none, make a transformation on the variable (Chapter 6).

   (b) Check for heteroscedasticity and autocorrelation (for time series data). If present, take appropriate action (Chapters 7 and 8).

   (c) Look for outliers, high leverage points, and influential points. If present, take appropriate action (Chapter 4).

4. Examine if additional variables can be dropped without compromising the integrity of the model. Examine if new variables are to be brought into the model (added variable plots, residual plus component plots) (Chapters 4 and 11). Repeat Step 3. Monitor the fitting process by examining the information

criteria (AIC or BIC). This is particularly relevant in examining non-nested models.

5. For the final fitted model, check variance inflation factors. Ensure satisfactory residual plots and no negative diagnostic messages (Chapters 3, 5, 6, and 9). If need be, repeat Step 4.

6. Attempt should then be made to validate the fitted model. When the amount of data is large, the model may be fitted by part of the data and validated by the remainder of the data. Resampling methods such as bootstrap, jackknife, and cross-validation are also possibilities, particularly when the amount of data available is not large [see Efron (1982) and Diaconis and Efron (1983)].

The steps we have described are, in practice, often not done sequentially but implemented synchronously. The process described is an iterative process and it may be necessary to recycle through the outlined steps several times to arrive at a satisfactory model. They enumerate the factors that must be considered for constructing a satisfactory model.

One important component that we have not included in our outlined steps is the subject matter knowledge of the analyst in the area in which the model is constructed. This knowledge should always be incorporated in the model-building process. Incorporation of this knowledge will often accelerate the process of arriving at a satisfactory model because it will help considerably in the appropriate choice of variables and corresponding transformations. After all is said and done, statistical model building is an art. The techniques that we have described are the tools by which this task can be attempted methodically.

11.16 BIBLIOGRAPHIC NOTES

There is a vast amount of literature on variable selection scattered in statistical journals. A very comprehensive review with an extensive bibliography may be found in Hocking (1976). A detailed treatment on variable selection with special emphasis on C_p statistic is given in the book by Daniel and Wood (1980). Refinements on the application of C_p statistic are given by Mallows (1973). The variable selection procedures are discussed in the book by Draper and Smith (1998). Use of ridge regression in connection with variable selection is discussed by Hoerl and Kennard (1970) and by McDonald and Schwing (1973).

EXERCISES

11.1 As we have seen in Section 11.14, the three noncollinear subsets of predictor variables below have emerged. Apply one or more variable selection methods to each subset and compare the resulting final models:

(a) The subset of eight variables: 1, 2, 3, 4, 5, 6, 9, and 14.

(b) The subset of 14 variables obtained after omitting variable 12.

Table 11.16 List of Variables for Data in Table 11.17

Variable	Definition
Y	Sale price of the house in thousands of dollars
X_1	Taxes (local, county, school) in thousands of dollars
X_2	Number of bathrooms
X_3	Lot size (in thousands of square feet)
X_4	Living space (in thousands of square feet)
X_5	Number of garage stalls
X_6	Number of rooms
X_7	Number of bedrooms
X_8	Age of of the home (years)
X_9	Number of fireplaces

(c) The subset of 14 variables obtained after omitting variable 13.

11.2 The estimated regression coefficients in Table 11.13 correspond to the standardized versions of the variables because they are computed using the correlation matrix of the response and predictor variables. Using the means and standard deviations of the variables in Table 11.11, write the estimated regresion equation in terms of the original variables (before centering and scaling).

11.3 In the Homicide data discussed in Section 11.12, we observed that when fitting the model in (11.11), the FS and BE methods give contradictory results. In fact, there are several other subsets in the data (not necessarily with three predictor variables) for which the FS and BE methods give contradictory results. Find one or more of these subsets.

11.4 Use the variable selection methods, as appropriate, to find one or more subsets of the predictor variables in Tables 11.7 and 11.8 that best account for the variability in the response variable H.

11.5 Property Valuation: Scientific mass appraisal is a technique in which linear regression methods applied to the problem of property valuation. The objective in scientific mass appraisal is to predict the sale price of a home from selected physical characteristics of the building and taxes (local, school, county) paid on the building. Twenty-four observations were obtained from *Multiple Listing* (Vol. 87) for Erie, PA, which is designated as Area 12 in the directory. These data (Table 11.17) were originally presented by Narula and Wellington (1977). The list of variables are given in Table 11.16.

Answer the following questions, in each case justifying your answer by appropriate analyses.

Table 11.17 Building Characteristics and Sales Price

Row	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Y
1	4.918	1.000	3.472	0.998	1.0	7	4	42	0	25.90
2	5.021	1.000	3.531	1.500	2.0	7	4	62	0	29.50
3	4.543	1.000	2.275	1.175	1.0	6	3	40	0	27.90
4	4.557	1.000	4.050	1.232	1.0	6	3	54	0	25.90
5	5.060	1.000	4.455	1.121	1.0	6	3	42	0	29.90
6	3.891	1.000	4.455	0.988	1.0	6	3	56	0	29.90
7	5.898	1.000	5.850	1.240	1.0	7	3	51	1	30.90
8	5.604	1.000	9.520	1.501	0.0	6	3	32	0	28.90
9	5.828	1.000	6.435	1.225	2.0	6	3	32	0	35.90
10	5.300	1.000	4.988	1.552	1.0	6	3	30	0	31.50
11	6.271	1.000	5.520	0.975	1.0	5	2	30	0	31.00
12	5.959	1.000	6.666	1.121	2.0	6	3	32	0	30.90
13	5.050	1.000	5.000	1.020	0.0	5	2	46	1	30.00
14	8.246	1.500	5.150	1.664	2.0	8	4	50	0	36.90
15	6.697	1.500	6.902	1.488	1.5	7	3	22	1	41.90
16	7.784	1.500	7.102	1.376	1.0	6	3	17	0	40.50
17	9.038	1.000	7.800	1.500	1.5	7	3	23	0	43.90
18	5.989	1.000	5.520	1.256	2.0	6	3	40	1	37.90
19	7.542	1.500	5.000	1.690	1.0	6	3	22	0	37.90
20	8.795	1.500	9.890	1.820	2.0	8	4	50	1	44.50
21	6.083	1.500	6.727	1.652	1.0	6	3	44	0	37.90
22	8.361	1.500	9.150	1.777	2.0	8	4	48	1	38.90
23	8.140	1.000	8.000	1.504	2.0	7	3	3	0	36.90
24	9.142	1.500	7.326	1.831	1.5	8	4	31	0	45.80

(a) In a fitted regression model that relates the sale price to taxes and building characteristics, would you include all the variables?

(b) A veteran real estate agent has suggested that local taxes, number of rooms, and age of the house would adequately describe the sale price. Do you agree?

(c) A real estate expert who was brought into the project reasoned as follows: The selling price of a home is determined by its desirability and this is certainly a function of the physical characteristic of the building. This overall assessment is reflected in the local taxes paid by the homeowner; consequently, the best predictor of sale price is the local taxes. The building characteristics are therefore redundant in a regression equation which includes local taxes. An equation that relates sale price solely to local taxes would be adequate. Examine this assertion by examining several models. Do you agree? Present what you consider to be the most adequate model or models for predicting sale price of homes in Erie, PA.

11.6 Refer to the Gasoline Consumption data in Tables 9.18 and 9.19.

(a) Would you include all the variables to predict the gasoline consumption of the cars? Explain, giving reasons.

(b) Six alternative models have been suggested:
   (a) Regress Y on X_1.
   (b) Regress Y on X_{10}.
   (c) Regress Y on X_1 and X_{10}.
   (d) Regress Y on X_2 and X_{10}.
   (e) Regress Y on X_8 and X_{10}.
   (f) Regress Y on X_8 and X_5, and X_{10}.

   Among these regression models, which would you choose to predict the gasoline consumption of automobiles? Can you suggest a better model?

(c) Plot Y against X_1, X_2, X_8, and X_{10} (one at a time). Do the plots suggest that the relationship between Y and the 11 predictor variables may not be linear?

(d) The gasoline consumption was determined by driving each car with the same load over the same track (a road length of about 123 miles). Instead of using Y (miles per gallon), it was suggested that we consider a new variable, $W = 100/Y$ (gallons per hundred miles). Plot W against X_1, X_2, X_8, and X_{10} and examine if the relationship between W and the 11 predictor variables is more linear than that between Y and the 11 predictor variables.

(e) Repeat Part (b) using W in place of Y. What are your conclusions?

(f) Regress Y on X_{13}, where $X_{13} = X_8/X_{10}$.

(g) Write a brief report describing your findings. Make a recommendation on the model to be used for predicting gasoline consumption of cars.

11.7 Refer to the Presidential Election Data in Table 5.17 and, as in Exercise 9.3, consider fitting a model relating V to all the variables (including a time trend representing year of election) plus as many interaction terms involving two or three variables as you possibly can.

(a) Starting with the model in Exercise 9.3(a). Apply two or more variable selection methods to choose the best model or models that might be expected to perform best in predicting future presidential elections.

(b) Repeat the above exercise starting with the model in Exercise 9.3(d).

(c) Which one of the models obtained above would you prefer?

(d) Use your chosen model to predict the proportion of votes expected to be obtained by a presidential candidate in United States presidential elections in the years 2000, 2004, and 2008.

(e) Which one of the above three predictions would you expect to be more accurate than the other two? Explain.

(f) The result of the 2000 presidential election was not known at the time this edition went to press. If you happen to be reading this book after the election of the year 2000 and beyond, were your predictions in Exercise correct?

11.8 Cigarette Consumption Data: Consider the Cigarette Consumption data described in Exercise 3.14 and given in Table 3.17. The organization wanted to construct a regression equation that relates statewide cigarette consumption (per capita basis) to various socioeconomic and demographic variables, and to determine whether these variables were useful in predicting the consumption of cigarettes.

(a) Construct a linear regression model that explains the per capita sale of cigarettes in a given state. In your analysis, pay particular attention to outliers. See if the deletion of an outlier affects your findings. Look at residual plots before deciding on a final model. You need not include all the variables in the model if your analysis indicates otherwise. Your objective should be to find the smallest number of variables that describes the state sale of cigarettes meaningfully and adequately.

(b) Write a report describing your findings.

Appendix: Effects of Incorrect Model Specifications

In this Appendix we discuss the effects of an incorrect model specification on the estimates of the regression coefficients and predicted values using matrix notation. Define the following matrix and vectors:

$$\mathbf{X} = \left[\begin{array}{cccc|ccc} x_{10} & x_{11} & \cdots & x_{1p} & x_{1(p+1)} & \cdots & x_{1q} \\ x_{20} & x_{21} & \cdots & x_{2p} & x_{2(p+1)} & \cdots & x_{2q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} & x_{n(p+1)} & \cdots & x_{nq} \end{array}\right], \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \\ \hline \beta_{p+1} \\ \vdots \\ \beta_q \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where $x_{i0} = 1$ for $i = 1, \ldots, n$. The matrix $\mathbf{X}$, which has n rows and $(q+1)$ columns, is partitioned into two submatrices $\mathbf{X}_p$ and $\mathbf{X}_r$, of dimensions $(n \times (p+1))$ and $(n \times r)$, where $r = q - p$. The vector $\boldsymbol{\beta}$ is similarly partitioned into $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_r$, which have $(p+1)$ and r components, respectively.

The full linear model containing all q variables is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}, \tag{A.1}$$

where ε_i's are independently normally distributed errors with zero means and unit variance.

The linear model containing only p variables (i.e., an equation with $(p+1)$ terms) is

$$\mathbf{Y} = \mathbf{X}_p\boldsymbol{\beta}_p + \boldsymbol{\varepsilon}. \tag{A.2}$$

Let us denote the least squares estimate of $\boldsymbol{\beta}$ obtained from the full model (A.1) by $\hat{\boldsymbol{\beta}}^*$, where

$$\hat{\boldsymbol{\beta}}^* = \begin{pmatrix} \hat{\boldsymbol{\beta}}_p^* \\ \hat{\boldsymbol{\beta}}_r^* \end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

The estimate $\hat{\boldsymbol{\beta}}_p$ of $\boldsymbol{\beta}_p$ obtained from the subset model (A.2) is given by

$$\hat{\boldsymbol{\beta}}_p = (\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{X}_p^T\mathbf{Y}.$$

Let $\hat{\sigma}_q^2$ and $\hat{\sigma}_p^2$ denote the estimates of σ^2 obtained from (A.1) and (A.2), respectively. Then it follows that

$$\hat{\sigma}_q^2 = \frac{\mathbf{Y}^T\mathbf{Y} - \hat{\boldsymbol{\beta}}^{*T}\mathbf{X}^T\mathbf{Y}}{n-q-1}$$

and

$$\hat{\sigma}_p^2 = \frac{\mathbf{Y}^T\mathbf{Y} - \hat{\boldsymbol{\beta}}_p^T\mathbf{X}_p^T\mathbf{Y}}{n-p-1}.$$

It is known from standard theory that $\hat{\boldsymbol{\beta}}^*$ and $\hat{\sigma}_q^2$ are unbiased estimates of $\boldsymbol{\beta}$ and σ^2. It can be shown that

$$E(\hat{\boldsymbol{\beta}}_p) = \boldsymbol{\beta}_p + \mathbf{A}\boldsymbol{\beta}_r,$$

where

$$\mathbf{A} = (\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{X}_p^T\mathbf{X}_r.$$

Further,

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}_p) &= (\mathbf{X}_p^T\mathbf{X}_p)^{-1}\,\sigma^2, \\ Var(\hat{\boldsymbol{\beta}}^*) &= (\mathbf{X}^T\mathbf{X})^{-1}\,\sigma^2, \end{aligned}$$

and

$$\text{MSE}(\hat{\boldsymbol{\beta}}_p) = (\mathbf{X}_p^T\mathbf{X}_p)^{-1}\sigma^2 + \mathbf{A}\boldsymbol{\beta}_r\boldsymbol{\beta}_r^T\mathbf{A}^T.$$

We can summarize the properties of $\hat{\boldsymbol{\beta}}_p$ and $\hat{\boldsymbol{\beta}}_p^*$ as follows:

1. $\hat{\boldsymbol{\beta}}_p$ is a biased estimate of $\boldsymbol{\beta}_p$ unless (1) $\boldsymbol{\beta}_r = 0$ or (2) $\mathbf{X}_p^T\mathbf{X}_r = 0$.

2. The matrix $Var(\hat{\boldsymbol{\beta}}^*) - Var(\hat{\boldsymbol{\beta}}_p)$ is positive semidefinite; that is, variances of the least squares estimates of regression coefficients obtained from the full model are larger than the corresponding variances of the estimates obtained from the subset model. In other words, the deletion of variables always results in smaller variances for the estimates of the regression coefficients of the remaining variables.

3. If the matrix $Var(\hat{\boldsymbol{\beta}}_r^*) - \boldsymbol{\beta}_r\boldsymbol{\beta}_r^T$ is positive semidefinite, then the matrix $Var(\hat{\boldsymbol{\beta}}_p^*) - \text{MSE}(\hat{\boldsymbol{\beta}}_p)$ is positive semidefinite. This means that the least squares estimates of regression coefficients obtained from the subset model have smaller mean square error than estimates obtained from the full model when the variables deleted have regression coefficients that are smaller than the standard deviation of the estimates of the coefficients.

4. $\hat{\sigma}_p^2$ is generally biased upward as an estimate of σ^2.

To see the effect of model misspecification on prediction, let us examine the prediction corresponding to an observation, say $\mathbf{x}^T = (\mathbf{x}_p^T : \mathbf{x}_r^T)$. Let $\hat{y}^*$ denote the predicted value corresponding to $\mathbf{x}^T$ when the full set of variables are used. Then $\hat{y}^* = \mathbf{x}^T\hat{\boldsymbol{\beta}}^*$ with mean $\mathbf{x}^T\boldsymbol{\beta}$ and prediction variance $Var(\hat{y}^*)$:

$$Var(\hat{y}^*) = \sigma^2(1 + \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}).$$

On the other hand, if the subset model (A.2) is used, the estimated predicted value $\hat{y} = \mathbf{x}_p^T\hat{\boldsymbol{\beta}}_p$ with mean

$$E(\hat{y}) = \mathbf{x}_p^T\boldsymbol{\beta}_p + \mathbf{x}_p^T\mathbf{A}\boldsymbol{\beta}_r$$

and prediction variance

$$Var(\hat{y}) = \sigma^2(1 + \mathbf{x}_p^T(\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{x}_p).$$

The prediction mean square error is given by

$$\text{MSE}(\hat{y}) = \sigma^2(1 + \mathbf{x}_p^T(\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{x}_p) + (\mathbf{x}_p^T\mathbf{A}\boldsymbol{\beta}_r - \mathbf{x}_r^T\boldsymbol{\beta}_r)^2.$$

The properties of $\hat{y}^*$ and $\hat{y}$ can be summarized as follows:

1. $\hat{y}$ is biased unless $\mathbf{X}_p^T\mathbf{X}_r\boldsymbol{\beta}_r = 0$.
2. $Var(\hat{y}^*) \geq Var(\hat{y})$.
3. If the matrix $Var(\hat{\boldsymbol{\beta}}_r^*) - \boldsymbol{\beta}_r\boldsymbol{\beta}_r^T$ is positive semidefinite, then $Var(\hat{y}^*) \geq \text{MSE}(\hat{y})$.

The significance and interpretation of these results in the context of variable selection are given in the main body of the chapter.