Chapter 9

Building the Regression Model I: Model Selection and Validation

In earlier chapters, we considered how to fit simple and multiple regression models and how to make inferences from these models. In this chapter, we first present an overview of the model-building and model-validation process. Then we consider in more detail some special issues in the selection of the predictor variables for exploratory observational studies. We conclude the chapter with a detailed description of methods for validating regression models.

9.1 Overview of Model-Building Process

At the risk of oversimplifying, we present in Figure 9.1 a strategy for the building of a regression model. This strategy involves three or, sometimes, four phases:

1. Data collection and preparation
2. Reduction of explanatory or predictor variables (for exploratory observational studies)
3. Model refinement and selection
4. Model validation

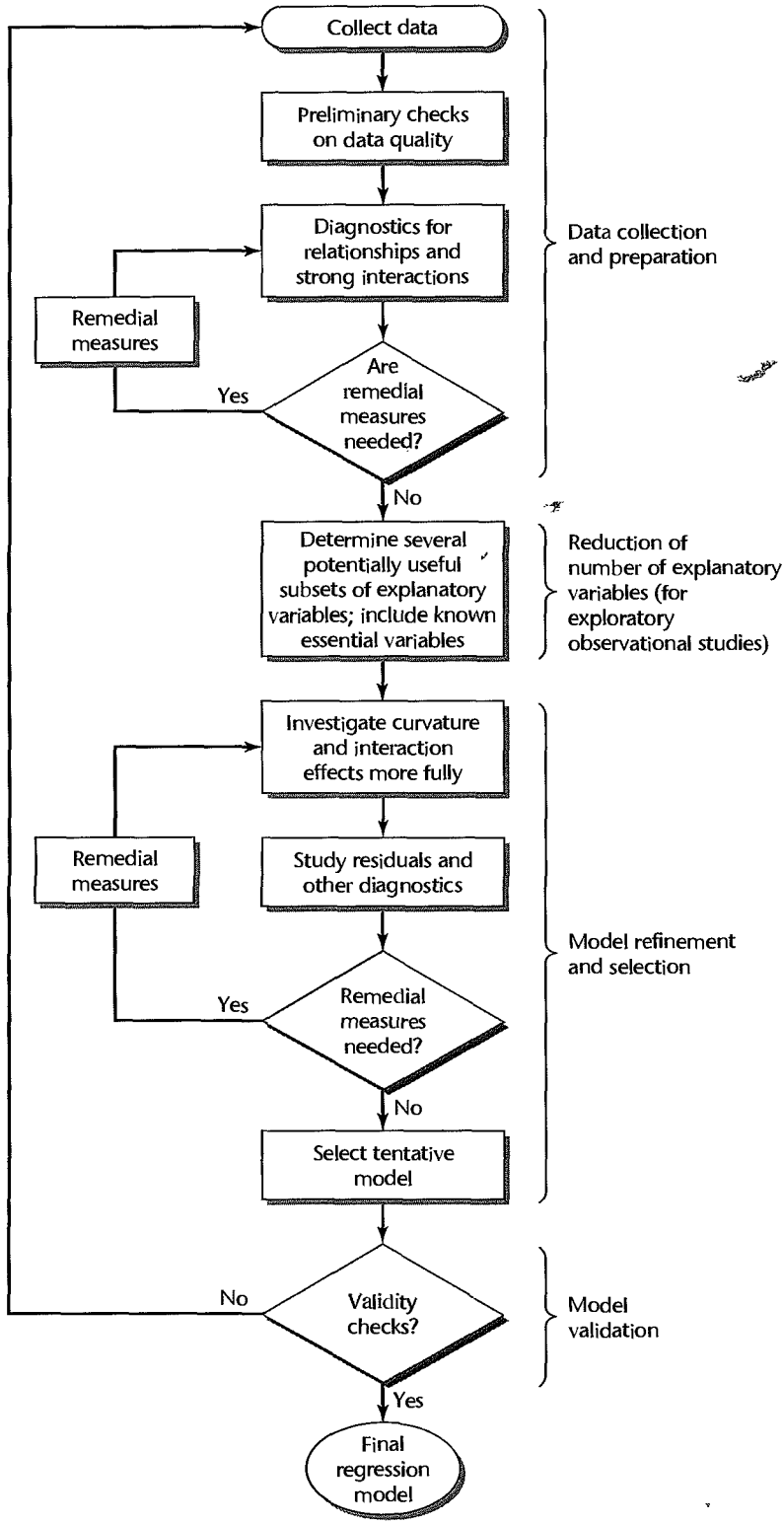
We consider each of these phases in turn.

Data Collection

The data collection requirements for building a regression model vary with the nature of the study. It is useful to distinguish four types of studies.

Controlled Experiments. In a controlled experiment, the experimenter controls the levels of the explanatory variables and assigns a treatment, consisting of a combination of levels of the explanatory variables, to each experimental unit and observes the response. For example, an experimenter studied the effects of the size of a graphic presentation and the time allowed for analysis of the accuracy with which the analysis of the presentation is carried out. Here, the response variable is a measure of the accuracy of the analysis, and the explanatory variables are the size of the graphic presentation and the time allowed. Junior

FIGURE 9.1
Strategy for Building a Regression Model.



executives were used as the experimental units. A treatment consisted of a particular combination of size of presentation and length of time allowed. In controlled experiments, the explanatory variables are often called *factors* or *control variables*.

The data collection requirements for controlled experiments are straightforward, though not necessarily simple. Observations for each experimental unit are needed on the response variable and on the level of each of the control variables used for that experimental unit. There may be difficult measurement and scaling problems for the response variable that are unique to the area of application.

Controlled Experiments with Covariates. Statistical design of experiments uses supplemental information, such as characteristics of the experimental units, in designing the experiment so as to reduce the variance of the experimental error terms in the regression model. Sometimes, however, it is not possible to incorporate this supplemental information into the design of the experiment. Instead, it may be possible for the experimenter to incorporate this information into the regression model and thereby reduce the error variance by including *uncontrolled variables* or *covariates* in the model.

In our previous example involving the accuracy of analysis of graphic presentations, the experimenter suspected that gender and number of years of education could affect the accuracy responses in important ways. Because of time constraints, the experimenter was able to use only a completely randomized design, which does not incorporate any supplemental information into the design. The experimenter therefore also collected data on two uncontrolled variables (gender and number of years of education of the junior executives) in case that use of these covariates in the regression model would make the analysis of the effects of the explanatory variables (size of graphic presentation, time allowed) on the accuracy response more precise.

Confirmatory Observational Studies. These studies, based on observational, not experimental, data, are intended to test (i.e., to confirm or not to confirm) hypotheses derived from previous studies or from hunches. For these studies, data are collected for explanatory variables that previous studies have shown to affect the response variable, as well as for the new variable or variables involved in the hypothesis. In this context, the explanatory variable(s) involved in the hypothesis are sometimes called the *primary variables*, and the explanatory variables that are included to reflect existing knowledge are called the *control variables* (*known risk factors* in epidemiology). The control variables here are not controlled as in an experimental study, but they are used to account for known influences on the response variable. For example, in an observational study of the effect of vitamin E supplements on the occurrence of a certain type of cancer, known risk factors, such as age, gender, and race, would be included as control variables and the amount of vitamin E supplements taken daily would be the primary explanatory variable. The response variable would be the occurrence of the particular type of cancer during the period under consideration. (The use of qualitative response variables in a regression model will be considered in Chapter 14.)

Data collection for confirmatory observational studies involves obtaining observations on the response variable, the control variables, and the primary explanatory variable(s). Here, as in controlled experiments, there may be important and complex problems of measurement, such as how to obtain reliable data on the amount of vitamin supplements taken daily.

Exploratory Observational Studies. In the social, behavioral, and health sciences, management, and other fields, it is often not possible to conduct controlled experiments.

Furthermore, adequate knowledge for conducting confirmatory observational studies may be lacking. As a result, many studies in these fields are exploratory observational studies where investigators search for explanatory variables that might be related to the response variable. To complicate matters further, any available theoretical models may involve explanatory variables that are not directly measurable, such as a family's future earnings over the next 10 years. Under these conditions, investigators are often forced to prospect for explanatory variables that could conceivably be related to the response variable under study. Obviously, such a set of potentially useful explanatory variables can be large. For example, a company's sales of portable dishwashers in a district may be affected by population size, per capita income, percent of population in urban areas, percent of population under 50 years of age, percent of families with children at home, etc., etc.!

After a lengthy list of potentially useful explanatory variables has been compiled, some of these variables can be quickly screened out. An explanatory variable (1) may not be fundamental to the problem, (2) may be subject to large measurement errors, and/or (3) may effectively duplicate another explanatory variable in the list. Explanatory variables that cannot be measured may either be deleted or replaced by proxy variables that are highly correlated with them.

The number of cases to be collected for an exploratory observational regression study depends on the size of the pool of potentially useful explanatory variables available at this stage. More cases are required when the pool is large than when it is small. A general rule of thumb states that there should be at least 6 to 10 cases for every variable in the pool. The actual data collection for the pool of potentially useful explanatory variables and for the response variable again may involve important issues of measurement, just as for the other types of studies.

Data Preparation

Once the data have been collected, edit checks should be performed and plots prepared to identify gross data errors as well as extreme outliers. Difficulties with data errors are especially prevalent in large data sets and should be corrected or resolved before the model building begins. Whenever possible, the investigator should carefully monitor and control the data collection process to reduce the likelihood of data errors.

Preliminary Model Investigation

Once the data have been properly edited, the formal modeling process can begin. A variety of diagnostics should be employed to identify (1) the functional forms in which the explanatory variables should enter the regression model and (2) important interactions that should be included in the model. Scatter plots and residual plots are useful for determining relationships and their strengths. Selected explanatory variables can be fitted in regression functions to explore relationships, possible strong interactions, and the need for transformations. Whenever possible, of course, one should also rely on the investigator's prior knowledge and expertise to suggest appropriate transformations and interactions to investigate. This is particularly important when the number of potentially useful explanatory variables is large. In this case, it may be very difficult to investigate all possible pairwise interactions, and prior knowledge should be used to identify the important ones. The diagnostic procedures explained in previous chapters and in Chapter 10 should be used as resources in this phase of model building.

Reduction of Explanatory Variables

Controlled Experiments. The reduction of explanatory variables in the model-building phase is usually not an important issue for controlled experiments. The experimenter has chosen the explanatory variables for investigation, and a regression model is to be developed that will enable the investigator to study the effects of these variables on the response variable. After the model has been developed, including the use of appropriate functional forms for the variables and the inclusion of important interaction terms, the inferential procedures considered in previous chapters will be used to determine whether the explanatory variables have effects on the response variable and, if so, the nature and magnitude of the effects.

Controlled Experiments with Covariates. In studies of controlled experiments with covariates, some reduction of the covariates may take place because investigators often cannot be sure in advance that the selected covariates will be helpful in reducing the error variance. For instance, the investigator in our graphic presentation example may wish to examine at this stage of the model-building process whether gender and number of years of education are related to the accuracy response, as had been anticipated. If not, the investigator would wish to drop them as not being helpful in reducing the model error variance and, therefore, in the analysis of the effects of the explanatory variables on the response variable. The number of covariates considered in controlled experiments is usually small, so no special problems are encountered in determining whether some or all of the covariates should be dropped from the regression model.

Confirmatory Observational Studies. Generally, no reduction of explanatory variables should take place in confirmatory observational studies. The control variables were chosen on the basis of prior knowledge and should be retained for comparison with earlier studies even if some of the control variables turn out not to lead to any error variance reduction in the study at hand. The primary variables are the ones whose influence on the response variable is to be examined and therefore need to be present in the model.

Exploratory Observational Studies. In exploratory observational studies, the number of explanatory variables that remain after the initial screening typically is still large. Further, many of these variables frequently will be highly intercorrelated. Hence, the investigator usually will wish to reduce the number of explanatory variables to be used in the final model. There are several reasons for this. A regression model with numerous explanatory variables may be difficult to maintain. Further, regression models with a limited number of explanatory variables are easier to work with and understand. Finally, the presence of many highly intercorrelated explanatory variables may substantially increase the sampling variation of the regression coefficients, detract from the model's descriptive abilities, increase the problem of roundoff errors (as noted in Chapter 7), and not improve, or even worsen, the model's predictive ability. An actual worsening of the model's predictive ability can occur when explanatory variables are kept in the regression model that are not related to the response variable, given the other explanatory variables in the model. In that case, the variances of the fitted values $\sigma^2\{\hat{Y}_i\}$ tend to become larger with the inclusion of the useless additional explanatory variables.

Hence, once the investigator has tentatively decided upon the functional form of the regression relations (whether given variables are to appear in linear form, quadratic form, etc.) and whether any interaction terms are to be included, the next step in many exploratory

observational studies is to identify a few "good" subsets of X variables for further intensive study. These subsets should include not only the potential explanatory variables in first-order form but also any needed quadratic and other curvature terms and any necessary interaction terms.

The identification of "good" subsets of potentially useful explanatory variables to be included in the final regression model and the determination of appropriate functional and interaction relations for these variables usually constitute some of the most difficult problems in regression analysis. Since the uses of regression models vary, no one subset of explanatory variables may always be "best." For instance, a descriptive use of a regression model typically will emphasize precise estimation of the regression coefficients, whereas a predictive use will focus on the prediction errors. Often, different subsets of the pool of potential explanatory variables will best serve these varying purposes. Even for a given purpose, it is often found that several subsets are about equally "good" according to a given criterion, and the choice among these "good" subsets needs to be made on the basis of additional considerations.

The choice of a few appropriate subsets of explanatory variables for final consideration in exploratory observational studies needs to be done with great care. Elimination of key explanatory variables can seriously damage the explanatory power of the model and lead to biased estimates of regression coefficients, mean responses, and predictions of new observations, as well as biased estimates of the error variance. The bias in these estimates is related to the fact that with observational data, the error terms in an underfitted regression model may reflect nonrandom effects of the explanatory variables not incorporated in the regression model. Important omitted explanatory variables are sometimes called *latent explanatory variables*.

On the other hand, if too many explanatory variables are included in the subset, then this overfitted model will often result in variances of estimated parameters that are larger than those for simpler models.

Another danger with observational data is that important explanatory variables may be observed only over narrow ranges. As a result, such important explanatory variables may be omitted just because they occur in the sample within a narrow range of values and therefore turn out to be statistically nonsignificant.

Another consideration in identifying subsets of explanatory variables is that these subsets need to be small enough so that maintenance costs are manageable and analysis is facilitated, yet large enough so that adequate description, control, or prediction is possible.

A variety of computerized approaches have been developed to assist the investigator in reducing the number of potential explanatory variables in an exploratory observational study when these variables are correlated among themselves. We present two of these approaches in this chapter. The first, which is practical for pools of explanatory variables that are small or moderate in size, considers all possible subsets of explanatory variables that can be developed from the pool of potential explanatory variables and identifies those subsets that are "good" according to a criterion specified by the investigator. The second approach employs automatic search procedures to arrive at a single subset of the explanatory variables. This approach is recommended primarily for reductions involving large pools of explanatory variables.

Even though computerized approaches can be very helpful in identifying appropriate subsets for detailed, final consideration, the process of developing a useful regression model must be pragmatic and needs to utilize large doses of subjective judgment. Explanatory

variables that are considered essential should be included in the regression model before any computerized assistance is sought. Further, computerized approaches that identify only a single subset of explanatory variables as "best" need to be supplemented so that additional subsets are also considered before the final regression model is decided upon.

Comments

1. All too often, unwary investigators will screen a set of explanatory variables by fitting the regression model containing the entire set of potential X variables and then simply dropping those for which the t^* statistic (7.25):

$$t_k^* = \frac{b_k}{s\{b_k\}}$$

has a small absolute value. As we know from Chapter 7, this procedure can lead to the dropping of important intercorrelated explanatory variables. Clearly, a good search procedure must be able to handle important intercorrelated explanatory variables in such a way that not all of them will be dropped.

2. Controlled experiments can usually avoid many of the problems in exploratory observational studies. For example, the effects of latent predictor variables are minimized by using randomization. In addition, adequate ranges of the explanatory variables can be selected and correlations among the explanatory variables can be eliminated by appropriate choices of their levels. ■

Model Refinement and Selection

At this stage in the model-building process, the tentative regression model, or the several "good" regression models in the case of exploratory observational studies, need to be checked in detail for curvature and interaction effects. Residual plots are helpful in deciding whether one model is to be preferred over another. In addition, the diagnostic checks to be described in Chapter 10 are useful for identifying influential outlying observations, multicollinearity, etc.

The selection of the ultimate regression model often depends greatly upon these diagnostic results. For example, one fitted model may be very much influenced by a single case, whereas another is not. Again, one fitted model may show correlations among the error terms, whereas another does not.

When repeat observations are available, formal tests for lack of fit can be made. In any case, a variety of residual plots and analyses can be employed to identify any lack of fit, outliers, and influential observations. For instance, residual plots against cross-product and/or power terms not included in the regression model can be useful in identifying ways in which the model fit can be improved further.

When an automatic selection procedure is utilized for an exploratory observational study and only a single model is identified as "best," other models should also be explored. One procedure is to use the number of explanatory variables in the model identified as "best" as an estimate of the number of explanatory variables needed in the regression model. Then the investigator explores and identifies other candidate models with approximately the same number of explanatory variables identified by the automatic procedure.

Eventually, after thorough checking and various remedial actions, such as transformations, the investigator narrows the number of competing models to one or just a few. At this point, it is good statistical practice to assess the validity of the remaining candidates through model validation studies. These methods can be used to help decide upon a final regression model, and to determine how well the model will perform in practice.

Model Validation

Model validity refers to the stability and reasonableness of the regression coefficients, the plausibility and usability of the regression function, and the ability to generalize inferences drawn from the regression analysis. Validation is a useful and necessary part of the model-building process. Several methods of assessing model validity will be described in Section 9.6.

9.2 Surgical Unit Example

With the completion of this overview of the model-building process for a regression study, we next present an example that will be used to illustrate all stages of this process as they are taken up in this and the following two chapters. A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 108 patients was available for analysis. From each patient record, the following information was extracted from the preoperation evaluation:

X_1	blood clotting score
X_2	prognostic index
X_3	enzyme function test score
X_4	liver function test score
X_5	age, in years
X_6	indicator variable for gender (0 = male, 1 = female)
X_7 and X_8	indicator variables for history of alcohol use:

Alcohol Use	X_7	X_8
None	0	0
Moderate	1	0
Severe	0	1

These constitute the pool of potential explanatory or predictor variables for a predictive regression model. The response variable is survival time, which was ascertained in a follow-up study. A portion of the data on the potential predictor variables and the response variable is presented in Table 9.1. These data have already been screened and properly edited for errors.

TABLE 9.1 Potential Predictor Variables and Response Variable—Surgical Unit Example.

Case Number i	Blood-Clotting Score X_{i1}	Prognostic Index X_{i2}	Enzyme Test X_{i3}	Liver Test X_{i4}	Age X_{i5}	Gender X_{i6}	Alc. Use: Mod. X_{i7}	Alc. Use: Heavy X_{i8}	Survival Time Y_i	$Y_i' = \ln Y_i$
1	6.7	62	81	2.59	50	0	1	0	695	6.544
2	5.1	59	66	1.70	39	0	0	0	403	5.999
3	7.4	57	83	2.16	55	0	0	0	710	6.565
...	...	...	...	...	...	...	...	...	...	...
52	6.4	85	40	1.21	58	0	0	1	579	6.361
53	6.4	59	85	2.33	63	0	1	0	550	6.310
54	8.8	78	72	3.20	56	0	0	0	651	6.478

To illustrate the model-building procedures discussed in this and the next section, we will use only the first four explanatory variables. By limiting the number of potential explanatory variables, we can explain the procedures without overwhelming the reader with masses of computer printouts. We will also use only the first 54 of the 108 patients.

Since the pool of predictor variables is small, a reasonably full exploration of relationships and of possible strong interaction effects is possible at this stage of data preparation. Stem-and-leaf plots were prepared for each of the predictor variables (not shown). These highlighted several cases as outlying with respect to the explanatory variables. The investigator was thereby alerted to examine later the influence of these cases. A scatter plot matrix and the correlation matrix were also obtained (not shown).

A first-order regression model based on all predictor variables was fitted to serve as a starting point. A plot of residuals against predicted values for this fitted model is shown in Figure 9.2a. The plot suggests that both curvature and nonconstant error variance are apparent. In addition, some departure from normality is suggested by the normal probability plot of residuals in Figure 9.2b.

To make the distribution of the error terms more nearly normal and to see if the same transformation would also reduce the apparent curvature, the investigator examined the

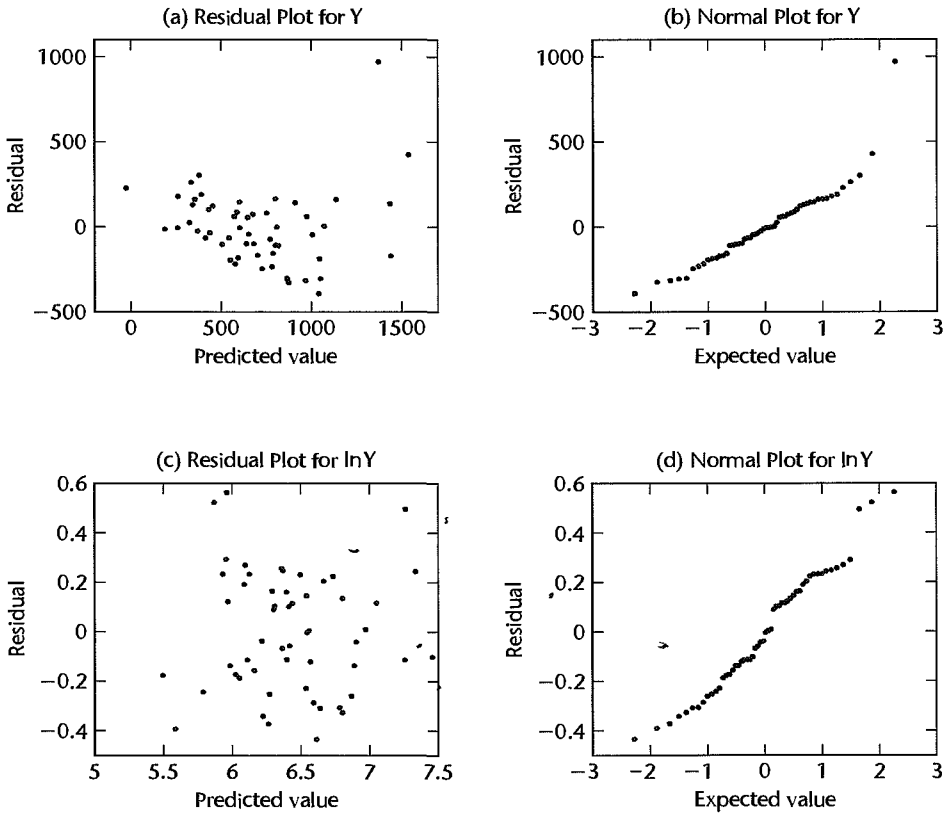


FIGURE 9.2 **Some Preliminary Residual Plots—Surgical Unit Example.**

logarithmic transformation $Y' = \ln Y$. Data for the transformed response variable are also given in Table 9.1. Figure 9.2c shows a plot of residuals against fitted values when Y' is regressed on all four predictor variables in a first-order model; also the normal probability plot of residuals for the transformed data shows that the distribution of the error terms is more nearly normal.

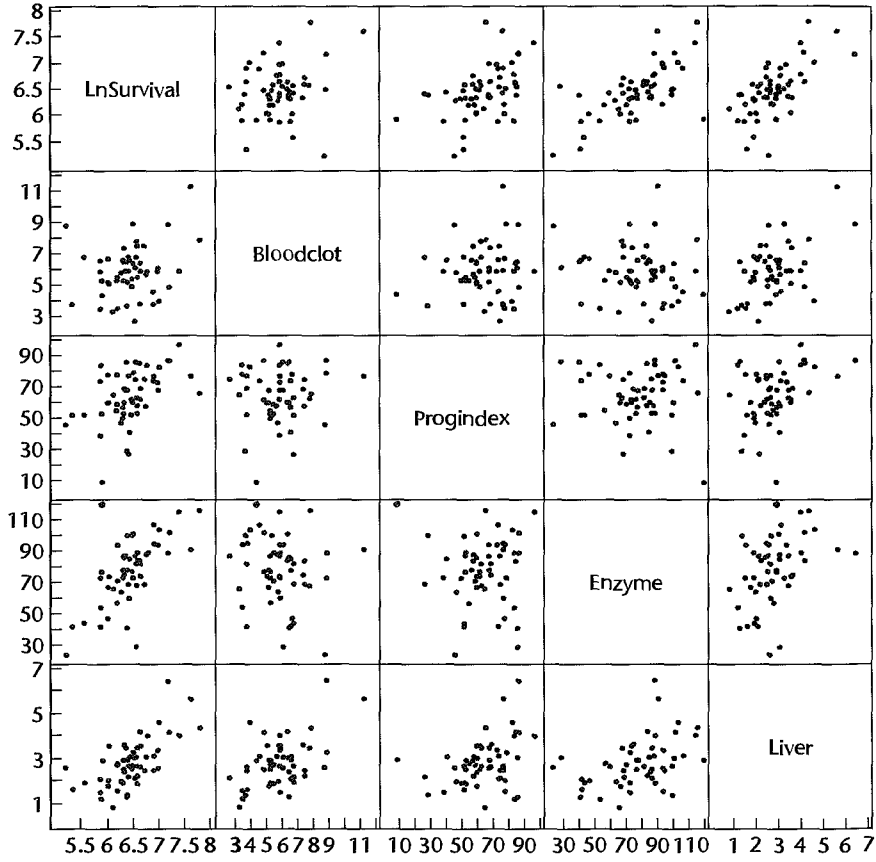
The investigator also obtained a scatter plot matrix and the correlation matrix with the transformed Y variable; these are presented in Figure 9.3. In addition, various scatter and

FIGURE 9.3 JMP Scatter Plot Matrix and Correlation Matrix when Response Variable Is Y'—Surgical Unit Example.

Multivariate Correlations

	LnSurvival	Bloodclot	Progindex	Enzyme	Liver
LnSurvival	1.0000	0.2462	0.4699	0.6539	0.6493
Bloodclot	0.2462	1.0000	0.0901	−0.1496	0.5024
Progindex	0.4699	0.0901	1.0000	−0.0236	0.3690
Enzyme	0.6539	−0.1496	−0.0236	1.0000	0.4164
Liver	0.6493	0.5024	0.3690	0.4164	1.0000

Scatterplot Matrix



residual plots were obtained (not shown here). All of these plots indicate that each of the predictor variables is linearly associated with Y', with X_3 and X_4 showing the highest degrees of association and X_1 the lowest. The scatter plot matrix and the correlation matrix further show intercorrelations among the potential predictor variables. In particular, X_4 has moderately high pairwise correlations with X_1, X_2, and X_3.

On the basis of these analyses, the investigator concluded to use, at this stage of the model-building process, $Y' = \ln Y$ as the response variable, to represent the predictor variables in linear terms, and not to include any interaction terms. The next stage in the model-building process is to examine whether all of the potential predictor variables are needed or whether a subset of them is adequate. A number of useful measures have been developed to assess the adequacy of the various subsets. We now turn to a discussion of these measures.

9.3 Criteria for Model Selection

From any set of $p-1$ predictors, 2^{p-1} alternative models can be constructed. This calculation is based on the fact that each predictor can be either included or excluded from the model. For example, the $2^4 = 16$ different possible subset models that can be formed from the pool of four X variables in the surgical unit example are listed in Table 9.2. First, there is the regression model with no X variables, i.e., the model $Y_i = \beta_0 + \varepsilon_i$. Then there are the regression models with one X variable (X_1, X_2, X_3, X_4), with two X variables (X_1 and X_2, X_1 and X_3, X_1 and X_4, X_2 and X_3, X_2 and X_4, X_3 and X_4), and so on.

TABLE 9.2 SSE_p, R_p^2, $R_{a,p}^2$, C_p, AIC_p, SBC_p, and $PRESS_p$ **Values for All Possible Regression Models—Surgical Unit Example.**

X Variables in Model	(1) p	(2) SSE_p	(3) R_p^2	(4) $R_{a,p}^2$	(5) C_p	(6) AIC_p	(7) SBC_p	(8) $PRESS_p$
None	1	12.808	0.000	0.000	151.498	−75.703	−73.714	13.296
X_1	2	12.031	0.061	0.043	141.164	−77.079	−73.101	13.512
X_2	2	9.979	0.221	0.206	108.556	−87.178	−83.200	10.744
X_3	2	7.332	0.428	0.417	66.489	−103.827	−99.849	8.327
X_4	2	7.409	0.422	0.410	67.715	−103.262	−99.284	8.025
X_1, X_2	3	9.443	0.263	0.234	102.031	−88.162	−82.195	11.062
X_1, X_3	3	5.781	0.549	0.531	43.852	−114.658	−108.691	6.988
X_1, X_4	3	7.299	0.430	0.408	67.972	−102.067	−96.100	8.472
X_2, X_3	3	4.312	0.663	0.650	20.520	−130.483	−124.516	5.065
X_2, X_4	3	6.622	0.483	0.463	57.215	−107.324	−101.357	7.476
X_3, X_4	3	5.130	0.599	0.584	33.504	−121.113	−115.146	6.121
X_1, X_2, X_3	4	3.109	0.757	0.743	3.391	−146.161	−138.205	3.914
X_1, X_2, X_4	4	6.570	0.487	0.456	58.392	−105.748	−97.792	7.903
X_1, X_3, X_4	4	4.968	0.612	0.589	32.932	−120.844	−112.888	6.207
X_2, X_3, X_4	4	3.614	0.718	0.701	11.424	−138.023	−130.067	4.597
X_1, X_2, X_3, X_4	5	3.084	0.759	0.740	5.000	−144.590	−134.645	4.069

In most circumstances, it will be impossible for an analyst to make a detailed examination of all possible regression models. For instance, when there are 10 potential X variables in the pool, there would be $2^{10} = 1{,}024$ possible regression models. With the availability of high-speed computers and efficient algorithms, running all possible regression models for 10 potential X variables is not time consuming. Still, the sheer volume of 1,024 alternative models to examine carefully would be an overwhelming task for a data analyst.

Model selection procedures, also known as subset selection or variables selection procedures, have been developed to identify a small group of regression models that are "good" according to a specified criterion. A detailed examination can then be made of a limited number of the more promising or "candidate" models, leading to the selection of the final regression model to be employed. This limited number might consist of three to six "good" subsets according to the criteria specified, so the investigator can then carefully study these regression models for choosing the final model.

While many criteria for comparing the regression models have been developed, we will focus on six: R_p^2, $R_{a,p}^2$, C_p, AIC_p, SBC_p, and $PRESS_p$. Before doing so, we will need to develop some notation. We shall denote the number of potential X variables in the pool by $P - 1$. We assume throughout this chapter that all regression models contain an intercept term β_0. Hence, the regression function containing all potential X variables contains P parameters, and the function with no X variables contains one parameter (β_0).

The number of X variables in a subset will be denoted by $p - 1$, as always, so that there are p parameters in the regression function for this subset of X variables. Thus, we have:

$$1 \leq p \leq P \tag{9.1}$$

We will assume that the number of observations exceeds the maximum number of potential parameters:

$$n > P \tag{9.2}$$

and, indeed, it is highly desirable that n be substantially larger than P, as we noted earlier, so that sound results can be obtained.

R_p^2 or SSE_p Criterion

The R_p^2 criterion calls for the use of the coefficient of multiple determination R^2, defined in (6.40), in order to identify several "good" subsets of X variables—in other words, subsets for which R^2 is high. We show the number of parameters in the regression model as a subscript of R^2. Thus R_p^2 indicates that there are p parameters, or $p - 1$ X variables, in the regression function on which R_p^2 is based.

The R_p^2 criterion is equivalent to using the error sum of squares SSE_p as the criterion (we again show the number of parameters in the regression model as a subscript). With the SSE_p criterion, subsets for which SSE_p is small are considered "good." The equivalence of the R_p^2 and SSE_p criteria follows from (6.40):

$$R_p^2 = 1 - \frac{SSE_p}{SSTO} \tag{9.3}$$

Since the denominator $SSTO$ is constant for all possible regression models, R_p^2 varies inversely with SSE_p.

The R_p^2 criterion is not intended to identify the subsets that maximize this criterion. We know that R_p^2 can never decrease as additional X variables are included in the model. Hence, R_p^2 will be a maximum when all $P - 1$ potential X variables are included in the regression model. The intent in using the R_p^2 criterion is to find the point where adding more X variables is not worthwhile because it leads to a very small increase in R_p^2. Often, this point is reached when only a limited number of X variables is included in the regression model. Clearly, the determination of where diminishing returns set in is a judgmental one.

Example

Table 9.2 for the surgical unit example shows in columns 1 and 2 the number of parameters in the regression function and the error sum of squares for each possible regression model. In column 3 are given the R_p^2 values. The results were obtained from a series of computer runs. For instance, when X_4 is the only X variable in the regression model, we obtain:

$$R_2^2 = 1 - \frac{SSE(X_4)}{SSTO} = 1 - \frac{7.409}{12.808} = .422$$

Note that $SSTO = SSE_1 = 12.808$.

Figure 9.4a contains a plot of the R_p^2 values against p, the number of parameters in the regression model. The maximum R_p^2 value for the possible subsets each consisting of $p - 1$ predictor variables, denoted by $\max(R_p^2)$, appears at the top of the graph for each p. These points are connected by solid lines to show the impact of adding additional X variables. Figure 9.4a makes it clear that little increase in $\max(R_p^2)$ takes place after three X variables are included in the model. Hence, consideration of the subsets (X_1, X_2, X_3) for which $R_4^2 = .757$ (as shown in column 3 of Table 9.2) and (X_2, X_3, X_4) for which $R_4^2 = .718$ appears to be reasonable according to the R_p^2 criterion.

Note that variables X_3 and X_4, correlate most highly with the response variable, yet this pair does not appear together in the $\max(R_p^2)$ model for $p = 4$. This suggests that X_1, X_2, and X_3 contain much of the information presented by X_4. Note also that the coefficient of multiple determination associated with subset (X_2, X_3, X_4), $R_4^2 = .718$, is somewhat smaller than $R_4^2 = .757$ for subset (X_1, X_2, X_3).

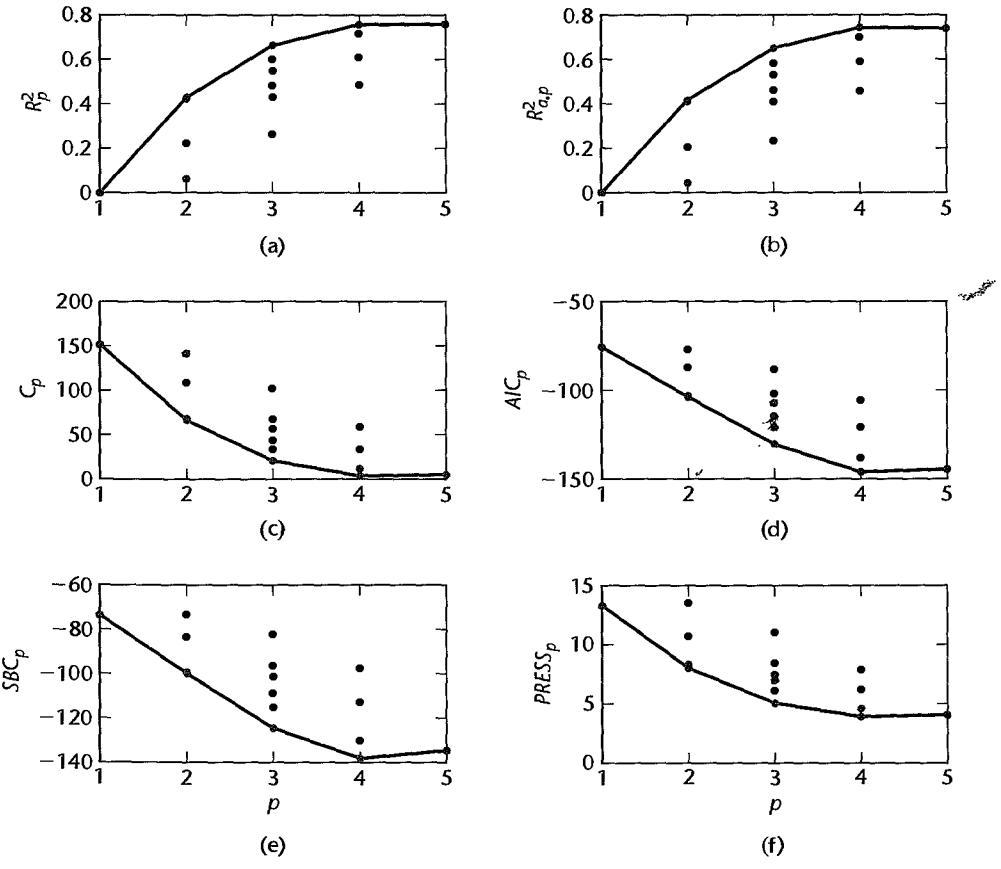
$R_{a,p}^2$ or MSE_p Criterion

Since R_p^2 does not take account of the number of parameters in the regression model and since $\max(R_p^2)$ can never decrease as p increases, the adjusted coefficient of multiple determination $R_{a,p}^2$ in (6.42) has been suggested as an alternative criterion:

$$R_{a,p}^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE_p}{SSTO} = 1 - \frac{\frac{MSE_p}{SSTO}}{n-1} \tag{9.4}$$

This coefficient takes the number of parameters in the regression model into account through the degrees of freedom. It can be seen from (9.4) that $R_{a,p}^2$ increases if and only if MSE_p decreases since $SSTO/(n-1)$ is fixed for the given Y observations. Hence, $R_{a,p}^2$ and MSE_p provide equivalent information. We shall consider here the criterion $R_{a,p}^2$, again showing the number of parameters in the regression model as a subscript of the criterion. The largest $R_{a,p}^2$ for a given number of parameters in the model, $\max(R_{a,p}^2)$, can, indeed, decrease as p increases. This occurs when the increase in $\max(R_p^2)$ becomes so small that it is not

FIGURE 9.4 Plot of Variables Selection Criteria—Surgical Unit Example.



sufficient to offset the loss of an additional degree of freedom. Users of the $R_{a,p}^2$ criterion seek to find a few subsets for which $R_{a,p}^2$ is at the maximum or so close to the maximum that adding more variables is not worthwhile.

Example

The $R_{a,p}^2$ values for all possible regression models for the surgical unit example are shown in Table 9.2, column 4. For instance, we have for the regression model containing only X_4:

$$R_{a,2}^2 = 1 - \left(\frac{n-1}{n-2}\right)\frac{SSE(X_4)}{SSTO} = 1 - \left(\frac{53}{52}\right)\frac{7.409}{12.808} = .410$$

Figure 9.4b contains the $R_{a,p}^2$ plot for the surgical unit example. We have again connected the $\max(R_{a,p}^2)$ values by solid lines. The story told by the $R_{a,p}^2$ plot in Figure 9.4b is very similar to that told by the R_p^2 plot in Figure 9.4a. Consideration of the subsets (X_1, X_2, X_3) and (X_2, X_3, X_4) appears to be reasonable according to the $R_{a,p}^2$ criterion. Notice that $R_{a,4}^2 = .743$ is maximized for subset (X_1, X_2, X_3), and that adding X_4 to this subset—thus using all four predictors—decreases the criterion slightly: $R_{a,5}^2 = .740$.

Mallows' C_p Criterion

This criterion is concerned with the *total mean squared error* of the n fitted values for each subset regression model. The mean squared error concept involves the total error in each fitted value:

$$\hat{Y}_i - \mu_i \tag{9.5}$$

where μ_i is the true mean response when the levels of the predictor variables X_k are those for the ith case. This total error is made up of a bias component and a random error component:

1. The bias component for the ith fitted value $\hat{Y}_i$, also called the model error component, is:

$$E\{\hat{Y}_i\} - \mu_i \tag{9.5a}$$

where $E\{\hat{Y}_i\}$ is the expectation of the ith fitted value for the given regression model. If the fitted model is not correct, $E\{\hat{Y}_i\}$ will differ from the true mean response μ_i and the difference represents the bias of the fitted model.

2. The random error component for $\hat{Y}_i$ is:

$$\hat{Y}_i - E\{\hat{Y}_i\} \tag{9.5b}$$

This component represents the deviation of the fitted value $\hat{Y}_i$ for the given sample from the expected value when the ith fitted value is obtained by fitting the same regression model to all possible samples.

The mean squared error for $\hat{Y}_i$ is defined as the expected value of the square of the total error in (9.5)—in other words, the expected value of:

$$(\hat{Y}_i - \mu_i)^2 = [(E\{\hat{Y}_i\} - \mu_i) + (\hat{Y}_i - E\{\hat{Y}_i\})]^2$$

It can be shown that this expected value is:

$$E\{\hat{Y}_i - \mu_i\}^2 = (E\{\hat{Y}_i\} - \mu_i)^2 + \sigma^2\{\hat{Y}_i\} \tag{9.6}$$

where $\sigma^2\{\hat{Y}_i\}$ is the variance of the fitted value $\hat{Y}_i$. We see from (9.6) that the mean squared error for the fitted value $\hat{Y}_i$ is the sum of the squared bias and the variance of $\hat{Y}_i$.

The total mean squared error for all n fitted values $\hat{Y}_i$ is the sum of the n individual mean squared errors in (9.6):

$$\sum_{i=1}^{n}[(E\{\hat{Y}_i\} - \mu_i)^2 + \sigma^2\{\hat{Y}_i\}] = \sum_{i=1}^{n}(E\{\hat{Y}_i\} - \mu_i)^2 + \sum_{i=1}^{n}\sigma^2\{\hat{Y}_i\} \tag{9.7}$$

The criterion measure, denoted by Γ_p, is simply the total mean squared error in (9.7) divided by σ^2, the true error variance:

$$\Gamma_p = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(E\{\hat{Y}_i\} - \mu_i)^2 + \sum_{i=1}^{n}\sigma^2\{\hat{Y}_i\}\right] \tag{9.8}$$

The model which includes all $P-1$ potential X variables is assumed to have been carefully chosen so that $MSE(X_1, \ldots, X_{P-1})$ is an unbiased estimator of σ^2. It can then be shown that an estimator of Γ_p is C_p:

$$C_p = \frac{SSE_p}{MSE(X_1, \ldots, X_{P-1})} - (n - 2p) \tag{9.9}$$

where SSE_p is the error sum of squares for the fitted subset regression model with p parameters (i.e., with $p-1$ X variables).

When there is no bias in the regression model with $p-1$ X variables so that $E\{\hat{Y}_i\} \equiv \mu_i$, the expected value of C_p is approximately p:

$$E\{C_p\} \approx p \qquad \text{when } E\{\hat{Y}_i\} \equiv \mu_i \tag{9.10}$$

Thus, when the C_p values for all possible regression models are plotted against p, those models with little bias will tend to fall near the line $C_p = p$. Models with substantial bias will tend to fall considerably above this line. C_p values below the line $C_p = p$ are interpreted as showing no bias, being below the line due to sampling error. The C_p value for the regression model containing all $P-1$ X variables is, by definition, P. The C_p measure assumes that $MSE(X_1, \ldots, X_{P-1})$ is an unbiased estimator of σ^2, which is equivalent to assuming that this model contains no bias.

In using the C_p criterion, we seek to identify subsets of X variables for which (1) the C_p value is small and (2) the C_p value is near p. Subsets with small C_p values have a small total mean squared error, and when the C_p value is also near p, the bias of the regression model is small. It may sometimes occur that the regression model based on a subset of X variables with a small C_p value involves substantial bias. In that case, one may at times prefer a regression model based on a somewhat larger subset of X variables for which the C_p value is only slightly larger but which does not involve a substantial bias component. Reference 9.1 contains extended discussions of applications of the C_p criterion.

Example

Table 9.2, column 5, contains the C_p values for all possible regression models for the surgical unit example. For instance, when X_4 is the only X variable in the regression model, the C_p value is:

$$C_2 = \frac{SSE(X_4)}{\dfrac{SSE(X_1, X_2, X_3, X_4)}{n-5}} - [n - 2(2)]$$

$$= \frac{7.409}{\dfrac{3.084}{49}} - [54 - 2(2)] = 67.715$$

The C_p values for all possible regression models are plotted in Figure 9.4c. We find that C_p is minimized for subset (X_1, X_2, X_3). Notice that $C_p = 3.391 < p = 4$ for this model, indicating little or no bias in the regression model.

Note that use of all potential X variables (X_1, X_2, X_3, X_4) results in a C_p value of exactly P, as expected; here, $C_5 = 5.00$. Also note that use of subset (X_2, X_3, X_4) with C_p value $C_4 = 11.424$ would be poor because of the substantial bias with this model. Thus, the C_p criterion suggests only one subset (X_1, X_2, X_3) for the surgical unit example.

Comments

1. Effective use of the C_p criterion requires careful development of the pool of $P-1$ potential X variables, with the predictor variables expressed in appropriate form (linear, quadratic, transformed), and important interactions included, so that $MSE(X_1, \ldots, X_{P-1})$ provides an unbiased estimate of the error variance σ^2.
2. The C_p criterion places major emphasis on the fit of the subset model for the n sample observations. At times, a modification of the C_p criterion that emphasizes new observations to be predicted may be preferable.
3. To see why C_p as defined in (9.9) is an estimator of Γ_p, we need to utilize two results that we shall simply state. First, it can be shown that:

$$\sum_{i=1}^{n} \sigma^2\{\hat{Y}_i\} = p\sigma^2 \tag{9.11}$$

Thus, the total random error of the n fitted values $\hat{Y}_i$ increases as the number of variables in the regression model increases.

Further, it can be shown that:

$$E\{SSE_p\} = \sum (E\{\hat{Y}_i\} - \mu_i)^2 + (n-p)\sigma^2 \tag{9.12}$$

Hence, Γ_p in (9.8) can be expressed as follows:

$$\begin{aligned} \Gamma_p &= \frac{1}{\sigma^2}[E\{SSE_p\} - (n-p)\sigma^2 + p\sigma^2] \\ &= \frac{E\{SSE_p\}}{\sigma^2} - (n-2p) \end{aligned} \tag{9.13}$$

Replacing $E\{SSE_p\}$ by the estimator SSE_p and using $MSE(X_1, \ldots, X_{P-1})$ as an estimator of σ^2 yields C_p in (9.9).

4. To show that the C_p value for the regression model containing all $P-1$ X variables is P, we substitute in (9.9), as follows:

$$\begin{aligned} C_p &= \frac{SSE(X_1, \ldots, X_{P-1})}{\frac{SSE(X_1, \ldots, X_{P-1})}{n-P}} - (n-2P) \\ &= (n-P) - (n-2P) \\ &= P \end{aligned}$$

■

AIC_p and SBC_p Criteria

We have seen that both $R^2_{a,p}$ and C_p are model selection criteria that penalize models having large numbers of predictors. Two popular alternatives that also provide penalties for adding predictors are Akaike's information criterion (AIC_p) and Schwarz' Bayesian

criterion (SBC_p). We search for models that have small values of AIC_p or SBC_p, where these criteria are given by:

$$AIC_p = n \ln SSE_p - n \ln n + 2p \tag{9.14}$$

$$SBC_p = n \ln SSE_p - n \ln n + [\ln n]p \tag{9.15}$$

Notice that for both of these measures, the first term is $n \ln SSE_p$, which decreases as p increases. The second term is fixed (for a given sample size n), and the third term increases with the number of parameters, p. Models with small SSE_p will do well by these criteria, as long as the penalties—$2p$ for AIC_p and $[\ln n]p$ for SBC_p—are not too large. If $n \geq 8$ the penalty for SBC_p is larger than that for AIC_p; hence the SBC_p criterion tends to favor more parsimonious models.

Example

Table 9.2, columns 6 and 7, contains the AIC_p and SBC_p values for all possible regression models for the surgical unit example. When X_4 is the only X variable in the regression model, the AIC_p value is:

$$\begin{aligned} AIC_2 &= n \ln SSE_2 - n \ln n + 2p \\ &= 54 \ln 7.409 - 54 \ln 54 + 2(2) = -103.262 \end{aligned}$$

Similarly, the SBC_p value is:

$$\begin{aligned} SBC_2 &= n \ln SSE_2 - n \ln n + [\ln n]p \\ &= 54 \ln 7.409 - 54 \ln 54 + [\ln 54](2) = -99.284 \end{aligned}$$

The AIC_p and SBC_p values for all possible regression models are plotted in Figures 9.4d and e. We find that both of these criteria are minimized for subset (X_1, X_2, X_3).

$PRESS_p$ Criterion

The $PRESS_p$ (prediction sum of squares) criterion is a measure of how well the use of the fitted values for a subset model can predict the observed responses Y_i. The error sum of squares, $SSE = \sum(Y_i - \hat{Y}_i)^2$, is also such a measure. The *PRESS* measure differs from *SSE* in that each fitted value $\hat{Y}_i$ for the *PRESS* criterion is obtained by deleting the ith case from the data set, estimating the regression function for the subset model from the remaining $n - 1$ cases, and then using the fitted regression function to obtain the predicted value $\hat{Y}_{i(i)}$ for the ith case. We use the notation $\hat{Y}_{i(i)}$ now for the fitted value to indicate, by the first subscript i, that it is a predicted value for the ith case and, by the second subscript (i), that the ith case was omitted when the regression function was fitted.

The *PRESS* prediction error for the ith case then is:

$$Y_i - \hat{Y}_{i(i)} \tag{9.16}$$

and the $PRESS_p$ criterion is the sum of the squared prediction errors over all n cases:

$$PRESS_p = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(i)})^2 \tag{9.17}$$

Models with small $PRESS_p$ values are considered good candidate models. The reason is that when the prediction errors $Y_i - \hat{Y}_{i(i)}$ are small, so are the squared prediction errors and the sum of the squared prediction errors. Thus, models with small $PRESS_p$ values fit well in the sense of having small prediction errors.

$PRESS_p$ values can be calculated without requiring n separate regression runs, each time deleting one of the n cases. The relationship in (10.21) and (10.21a), to be explained in the next chapter, enables one to calculate all $\hat{Y}_{i(i)}$ values from a single regression run.

Example

Table 9.2, column 8, contains the $PRESS_p$ values for all possible regression models for the surgical unit example. The $PRESS_p$ values are plotted in Figure 9.4f. The message given by the $PRESS_p$ values in Table 9.2 and plot in Figure 9.4f is very similar to that told by the other criteria. We find that subsets (X_1, X_2, X_3) and (X_2, X_3, X_4) have small *PRESS* values; in fact, the set of all X variables (X_1, X_2, X_3, X_4) involves a slightly larger *PRESS* value than subset (X_1, X_2, X_3). The subset (X_2, X_3, X_4) involves a *PRESS* value of 4.597, which is moderately larger than the *PRESS* value of 3.914 for subset (X_1, X_2, X_3).

Comment

PRESS values can also be useful for model validation, as will be explained in Section 9.6. ■

9.4 Automatic Search Procedures for Model Selection

As noted in the previous section, the number of possible models, 2^{p-1}, grows rapidly with the number of predictors. Evaluating all of the possible alternatives can be a daunting endeavor. To simplify the task, a variety of automatic computer-search procedures have been developed. In this section, we will review the two most common approaches, namely "best" subsets regression and stepwise regression.

For the remainder of this chapter, we will employ the full set of eight predictors from the surgical unit data. Recall that these predictors are displayed in Table 9.1 on page 350 and described there as well.

"Best" Subsets Algorithms

Time-saving algorithms have been developed in which the best subsets according to a specified criterion are identified without requiring the fitting of all of the possible subset regression models. In fact, these algorithms require the calculation of only a small fraction of all possible regression models. For instance, if the C_p criterion is to be employed and the five best subsets according to this criterion are to be identified, these algorithms search for the five subsets of X variables with the smallest C_p values using much less computational effort than when all possible subsets are evaluated. These algorithms are called *"best" subsets algorithms*. Not only do these algorithms provide the best subsets according to the specified criterion, but they often also identify several "good" subsets for each possible number of X variables in the model to give the investigator additional helpful information in making the final selection of the subset of X variables to be employed in the regression model.

When the pool of potential X variables is very large, say greater than 30 or 40, even the "best" subset algorithms may require excessive computer time. Under these conditions, one of the stepwise regression procedures, described later in this section, may need to be employed to assist in the selection of X variables.

Example

For the eight predictors in the surgical unit example, we know there are $2^8 = 256$ possible models. Plots of the six model selection criteria discussed in this chapter are displayed in

FIGURE 9.5
Plot of Variable Selection Criteria with All Eight Predictors—Surgical Unit Example.

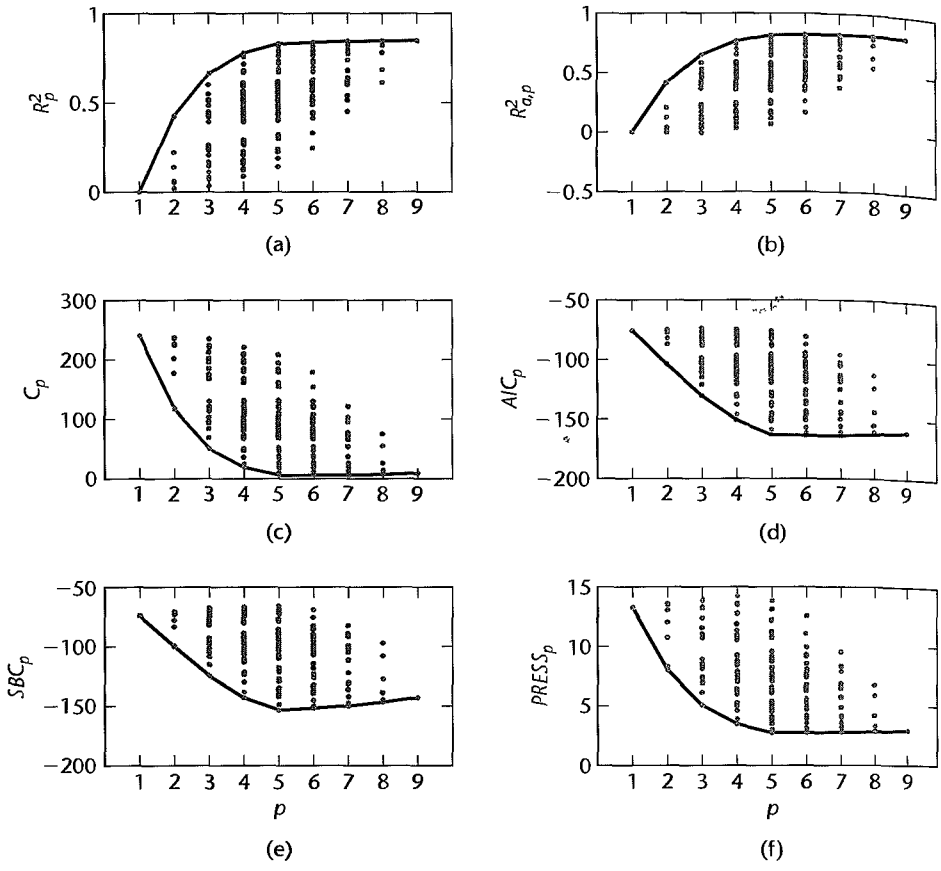


Figure 9.5. The best values of each criterion for each p have been connected with solid lines. These best values are also displayed in Table 9.3. The overall optimum criterion values have been underlined in each column of the table. Notice that the choice of a "best" model depends on the criterion. For example, a seven- or eight-parameter model is identified as best by the $R_{a,p}^2$ criterion (both have $\max(R_{a,p}^2) = .823$), a six-parameter model is identified by the C_p criterion ($\min(C_7) = 5.541$), and a seven-parameter model is identified by the AIC_p criterion ($\min(AIC_7) = -163.834$). As is frequently the case, the SBC_p criterion identifies a more parsimonious model as best. In this case both the SBC_p and $PRESS_p$ criteria point to five-parameter models ($\min(SBC_5) = -153.406$ and $\min(PRESS_5) = 2.738$). As previously emphasized, our objective at this stage is not to identify a single best model; we hope to identify a small set of promising models for further study.

Figure 9.6 contains, for the surgical unit example, MINITAB output for the "best" subsets algorithm. Here, we specified that the best two subsets be identified for each number of variables in the regression model. The MINITAB algorithm uses the R_p^2 criterion, but also shows for each of the "best" subsets the $R_{a,p}^2$, C_p, and $\sqrt{MSE_p}$ (labeled S) values. The right-most columns of the tabulation show the X variables in the subset. From the figure it is seen that the best subset, according to the $R_{a,p}^2$ criterion, is either the seven-parameter

TABLE 9.3 Best Variable-Selection Criterion Values—Surgical Unit Example.

p	(1) SSE_p	(2) R_p^2	(3) $R_{a,p}^2$	(4) C_p	(5) AIC_p	(6) SBC_p	(7) $PRESS_p$
1	12.808	0.000	0.000	240.452	−75.703	−73.714	13.296
2	7.332	0.428	0.417	117.409	−103.827	−99.849	8.025
3	4.312	0.663	0.650	50.472	−130.483	−124.516	5.065
4	2.843	0.778	0.765	18.914	−150.985	−143.029	3.469
5	2.179	0.830	0.816	5.751	−163.351	−153.406	2.738
6	2.082	0.837	0.821	5.541	−163.805	−151.871	2.739
7	2.005	0.843	0.823	5.787	−163.834	−149.911	2.772
8	1.972	0.846	0.823	7.029	−162.736	−146.824	2.809
9	1.971	0.846	0.819	9.000	−160.771	−142.870	2.931

FIGURE 9.6 MINITAB Output for "Best" Two Subsets for Each Subset Size—Surgical Unit Example.

```
Response is lnSurviv

                                                  B P           H
                                                  l r           i
                                                  o o E     G i s
                                                  o g n L   e s t
                                                  d i z i   n t h
                                                  c n y v A d m e
                                                  l d m e g e o a
Vars   R-Sq     R-Sq(adj)      C-p         S      o e e r e r d v

   1   42.8         41.7     117.4   0.37549        X
   1   42.2         41.0     119.2   0.37746          X
   2   66.3         65.0      50.5   0.29079      X X
   2   59.9         58.4      69.1   0.31715        X X
   3   77.8         76.5      18.9   0.23845      X X             X
   3   75.7         74.3      25.0   0.24934  X X X
   4   83.0         81.6       5.8   0.21087  X X X             X
   4   81.4         79.9      10.3   0.22023    X X X           X
   5   83.7         82.1       5.5   0.20827  X X X     X       X
   5   83.6         81.9       6.0   0.20931  X X X   X         X
   6   84.3         82.3       5.8   0.20655  X X X   X X       X
   6   83.9         81.9       7.0   0.20934  X X X     X X X   X
   7   84.6         82.3       7.0   0.20705  X X X   X X X X   X
   7   84.4         82.0       7.7   0.20867  X X X X X X       X
   8   84.6         81.9       9.0   0.20927  X X X X X X X X   X
```

model based on all predictors except Liver (X_4) and Histmod (history of moderate alcohol use—X_7), or the eight-parameter model based on all predictors except Liver (X_4). The $R_{a,p}^2$ criterion value for both of these models is .823.

The all-possible-regressions procedure leads to the identification of a small number of subsets that are "good" according to a specified criterion. In the surgical unit example, two of the four criteria—SBC_p and $PRESS_p$—pointed to models with 4 predictors, while the other criteria favored larger models. Consequently, one may wish at times to consider more than one criterion in evaluating possible subsets of X variables.

Once the investigator has identified a few "good" subsets for intensive examination, a final choice of the model variables must be made. This choice, as indicated by our model-building strategy in Figure 9.1, is aided by residual analyses (and other diagnostics to be covered in Chapter 10) and by the investigator's knowledge of the subject under study, and is finally confirmed through model validation studies.

Stepwise Regression Methods

In those occasional cases when the pool of potential X variables contains 30 to 40 or even more variables, use of a "best" subsets algorithm may not be feasible. An automatic search procedure that develops the "best" subset of X variables sequentially may then be helpful. The forward stepwise regression procedure is probably the most widely used of the automatic search methods. It was developed to economize on computational efforts, as compared with the various all-possible-regressions procedures. Essentially, this search method develops a sequence of regression models, at each step adding or deleting an X variable. The criterion for adding or deleting an X variable can be stated equivalently in terms of error sum of squares reduction, coefficient of partial correlation, t^* statistic, or F^* statistic.

An essential difference between stepwise procedures and the "best" subsets algorithm is that stepwise search procedures end with the identification of a *single* regression model as "best." With the "best" subsets algorithm, on the other hand, *several* regression models can be identified as "good" for final consideration. The identification of a single regression model as "best" by the stepwise procedures is a major weakness of these procedures. Experience has shown that each of the stepwise search procedures can sometimes err by identifying a suboptimal regression model as "best." In addition, the identification of a single regression model may hide the fact that several other regression models may also be "good." Finally, the "goodness" of a regression model can only be established by a thorough examination using a variety of diagnostics.

What then can we do on those occasions when the pool of potential X variables is very large and an automatic search procedure must be utilized? Basically, we should use the subset identified by the automatic search procedure as a starting point for searching for other "good" subsets. One possibility is to treat the number of X variables in the regression model identified by the automatic search procedure as being about the right subset size and then use the "best" subsets procedure for subsets of this and nearby sizes.

Forward Stepwise Regression

We shall describe the forward stepwise regression search algorithm in terms of the t^* statistics (2.17) and their associated P-values for the usual tests of regression parameters.

1. The stepwise regression routine first fits a simple linear regression model for each of the $P-1$ potential X variables. For each simple linear regression model, the t^* statistic (2.17) for testing whether or not the slope is zero is obtained:

$$t_k^* = \frac{b_k}{s\{b_k\}} \tag{9.18}$$

The X variable with the largest t^* value is the candidate for first addition. If this t^* value exceeds a predetermined level, or if the corresponding P-value is less than a predetermined α, the X variable is added. Otherwise, the program terminates with no X variable

considered sufficiently helpful to enter the regression model. Since the degrees of freedom associated with *MSE* vary depending on the number of X variables in the model, and since repeated tests on the same data are undertaken, fixed t^* limits for adding or deleting a variable have no precise probabilistic meaning. For this reason, software programs often favor the use of predetermined α-limits.

2. Assume X_7 is the variable entered at step 1. The stepwise regression routine now fits all regression models with two X variables, where X_7 is one of the pair. For each such regression model, the t^* test statistic corresponding to the newly added predictor X_k is obtained. This is the statistic for testing whether or not $\beta_k = 0$ when X_7 and X_k are the variables in the model. The X variable with the largest t^* value—or equivalently, the smallest P-value—is the candidate for addition at the second stage. If this t^* value exceeds a predetermined level (i.e., the P-value falls below a predetermined level), the second X variable is added. Otherwise, the program terminates.

3. Suppose X_3 is added at the second stage. Now the stepwise regression routine examines whether any of the other X variables already in the model should be dropped. For our illustration, there is at this stage only one other X variable in the model, X_7, so that only one t^* test statistic is obtained:

$$t_7^* = \frac{b_7}{s\{b_7\}} \tag{9.19}$$

At later stages, there would be a number of these t^* statistics, one for each of the variables in the model besides the one last added. The variable for which this t^* value is smallest (or equivalently the variable for which the P-value is largest) is the candidate for deletion. If this t^* value falls below—or the P-value exceeds—a predetermined limit, the variable is dropped from the model; otherwise, it is retained.

4. Suppose X_7 is retained so that both X_3 and X_7 are now in the model. The stepwise regression routine now examines which X variable is the next candidate for addition, then examines whether any of the variables already in the model should now be dropped, and so on until no further X variables can either be added or deleted, at which point the search terminates.

Note that the stepwise regression algorithm allows an X variable, brought into the model at an earlier stage, to be dropped subsequently if it is no longer helpful in conjunction with variables added at later stages.

Example

Figure 9.7 shows MINITAB computer printout for the forward stepwise regression procedure for the surgical unit example. The maximum acceptable α limit for adding a variable is 0.10 and the minimum acceptable α limit for removing a variable is 0.15, as shown at the top of Figure 9.7.

We now follow through the steps.

1. At the start of the stepwise search, no X variable is in the model so that the model to be fitted is $Y_i = \beta_0 + \varepsilon_i$. In step 1, the t^* statistics (9.18) and corresponding P-values are calculated for each potential X variable, and the predictor having the smallest P-value (largest t^* value) is chosen to enter the equation. We see that Enzyme (X_3) had the largest

FIGURE 9.7 MINITAB Forward Stepwise Regression Output—Surgical Unit Example.

```
Alpha-to-Enter: 0.1  Alpha-to-Remove: 0.15

Response is lnSurviv on  8 predictors, with N = 54

   Step          1        2        3        4
Constant     5.264    4.351    4.291    3.852

Enzyme      0.0151   0.0154   0.0145   0.0155
T-Value       6.23     8.19     9.33    11.07
P-Value      0.000    0.000    0.000    0.000

ProgInde             0.0141   0.0149   0.0142
T-Value                5.98     7.68     8.20
P-Value               0.000    0.000    0.000

Histheav                      0.429    0.353
T-Value                        5.08     4.57
P-Value                       0.000    0.000

Bloodclo                              0.073
T-Value                               3.86
P-Value                              0.000

S           0.375    0.291    0.238    0.211
R-Sq        42.76    66.33    77.80    82.99
R-Sq(adj)   41.66    65.01    76.47    81.60
C-p         117.4     50.5     18.9      5.8
```

test statistic:

$$t_3^* = \frac{b_3}{s\{b_3\}} = \frac{.015124}{.002427} = 6.23$$

The P-value for this test statistic is 0.000, which falls below the maximum acceptable α-to-enter value of 0.10; hence Enzyme (X_3) is added to the model.

2. At this stage, step 1 has been completed. The current regression model contains Enzyme (X_3), and the printout displays, near the top of the column labeled "Step 1," the regression coefficient for Enzyme (0.0151), the t^* value for this coefficient (6.23), and the corresponding P-value (0.000). At the bottom of column 1, a number of variables-selection criteria, including R_1^2 (42.76), $R_{a,1}^2$ (41.66), and C_1 (117.4) are also provided.

Next, all regression models containing X_3 and another X variable are fitted, and the t^* statistics calculated. They are now:

$$t_k^* = \sqrt{\frac{MSR(X_k|X_3)}{MSE(X_3, X_k)}}$$

Progindex (X_2) has the highest t^* value, and its P-value (0.000) falls below 0.10, so that X_2 now enters the model.

3. The column labeled Step 2 in Figure 9.7 summarizes the situation at this point. Enzyme and Progindex (X_3 and X_2) are now in the model, and information about this model is provided. At this point, a test whether Enzyme (X_3) should be dropped is undertaken, but because the P-value (0.000) corresponding to X_3 is not above 0.15, this variable is retained.

4. Next, all regression models containing X_2, X_3, and one of the remaining potential X variables are fitted. The appropriate t^* statistics now are:

$$t_k^* = \sqrt{\frac{MSR(X_k|X_2, X_3)}{MSE(X_2, X_3, X_k)}}$$

The predictor labeled Histheavy (X_8) had the largest t_k^* value (P-value = 0.000) and was next added to the model.

5. The column labeled Step 3 in Figure 9.7 summarizes the situation at this point. X_2, X_3, and X_8 are now in the model. Next, a test is undertaken to determine whether X_2 or X_3 should be dropped. Since both of the corresponding P-values are less than 0.15, neither predictor is dropped from the model.

6. At step 4 Bloodclot (X_1) is added, and no terms previously included were dropped. The right-most column of Figure 9.7 summarizes the addition of variable X_1 into the model containing variables X_2, X_3, and X_8. Next, a test is undertaken to determine whether either X_2, X_3, or X_8 should be dropped. Since all P-values are less than 0.15 (all are 0.000), all variables are retained.

7. Finally, the stepwise regression routine considers adding one of X_4, X_5, X_6, or X_7 to the model containing X_1, X_2, X_3, and X_8. In each case, the P-values are greater than 0.10 (not shown); therefore, no additional variables can be added to the model and the search process is terminated.

Thus, the stepwise search algorithm identifies (X_1, X_2, X_3, X_8) as the "best" subset of X variables. This model also happens to be the model identified by both the SBC_p and $PRESS_p$ criteria in our previous analyses based on an assessment of "best" subset selection.

Comments

1. The choice of α-to-enter and α-to-remove values essentially represents a balancing of opposing tendencies. Simulation studies have shown that for large pools of uncorrelated predictor variables that have been generated to be uncorrelated with the response variable, use of large or moderately large α-to-enter values as the entry criterion results in a procedure that is too liberal; that is, it allows too many predictor variables into the model. On the other hand, models produced by an automatic selection procedure with small α-to-enter values are often underspecified, resulting in σ^2 being badly overestimated and the procedure being too conservative (see, for example, References 9.2 and 9.3).

2. The maximum acceptable α-to-enter value should never be larger than the minimum acceptable α-to-remove value; otherwise, cycling is possible where a variable is continually entered and removed.

3. The order in which variables enter the regression model does not reflect their importance. At times, a variable may enter the model, only to be dropped at a later stage because it can be predicted well from the other predictors that have been subsequently added. ■

Other Stepwise Procedures

Other stepwise procedures are available to find a "best" subset of predictor variables. We mention two of these.

Forward Selection. The forward selection search procedure is a simplified version of forward stepwise regression, omitting the test whether a variable once entered into the model should be dropped.

Backward Elimination. The backward elimination search procedure is the opposite of forward selection. It begins with the model containing all potential X variables and identifies the one with the largest P-value. If the maximum P-value is greater than a predetermined limit, that X variable is dropped. The model with the remaining $P - 2$ X variables is then fitted, and the next candidate for dropping is identified. This process continues until no further X variables can be dropped. A stepwise modification can also be adapted that allows variables eliminated earlier to be added later; this modification is called the backward stepwise regression procedure.

Comment

For small and moderate numbers of variables in the pool of potential X variables, some statisticians argue for backward stepwise search over forward stepwise search (see Reference 9.4). A potential disadvantage of the forward stepwise approach is that the *MSE*—and hence $s\{b_k\}$—will tend to be inflated during the initial steps, because important predictors have been omitted. This in turn leads to t_k^* test statistics (9.18) that are too small. For the backward stepwise procedure, *MSE* values tend to be more nearly unbiased because important predictors are retained at each step. An argument in favor of the backward stepwise procedure can also be made in situations where it is useful as a first step to look at each X variable in the regression function adjusted for all the other X variables in the pool. ■

9.5 Some Final Comments on Automatic Model Selection Procedures

Our discussion of the major automatic selection procedures for identifying the "best" subset of X variables has focused on the main conceptual issues and not on options, variations, and refinements available with particular computer packages. It is essential that the specific features of the package employed be fully understood so that intelligent use of the package can be made. In some packages, there is an option for regression models through the origin. Some packages permit variables to be brought into the model and tested in pairs or other groupings instead of singly, to save computing time or for other reasons. Some packages, once a "best" regression model is identified, will fit all the possible regression models with the same number of variables and will develop information for each model so that a final choice can be made by the user. Some stepwise programs have options for forcing variables into the regression model; such variables are not removed even if their P-values become too large.

The diversity of these options and special features serves to emphasize a point made earlier: there is no unique way of searching for "good" subsets of X variables, and subjective elements must play an important role in the search process.

We have considered a number of important issues related to exploratory model building, but there are many others. (A good discussion of many of these issues may be found in Reference 9.5.) Most important for good model building is the recognition that no automatic search procedure will always find the "best" model, and that, indeed, there may exist several "good" regression models whose appropriateness for the purpose at hand needs to be investigated.

Judgment needs to play an important role in model building for exploratory studies. Some explanatory variables may be known to be more fundamental than others and therefore should be retained in the regression model if the primary purpose is to develop a good explanatory model. When a qualitative predictor variable is represented in the pool of potential X variables by a number of indicator variables (e.g., geographic region is represented by several indicator variables), it is often appropriate to keep these indicator variables together as a group to represent the qualitative variable, even if a subset containing only some of the indicator variables is "better" according to the criterion employed. Similarly, if second-order terms X_k^2 or interaction terms $X_k X_{k'}$ need to be present in a regression model, one would ordinarily wish to have the first-order terms in the model as representing the main effects.

The selection of a subset regression model for exploratory observational studies has been the subject of much recent research. Reference 9.5 provides information about many of these studies. New methods of identifying the "best" subset have been proposed, including methods based on deleting one case at a time and on bootstrapping. With the first method, the criterion is evaluated for identified subsets n times, each time with one case omitted, in order to select the "best" subset. With bootstrapping, repeated samples of cases are selected with replacement from the data set (alternatively, repeated samples of residuals from the model fitted to all X variables are selected with replacement to obtain observed Y values), and the criterion is evaluated for identified subsets in order to select the "best" subset. Research by Breiman and Spector (Ref. 9.7) has evaluated these methods from the standpoint of the closeness of the selected model to the true model and has found the two methods promising, the bootstrap method requiring larger data sets.

An important issue in exploratory model building that we have not yet considered is the bias in estimated regression coefficients and in estimated mean responses, as well as in their estimated standard deviations, that may result when the coefficients and error mean square for the finally selected regression model are estimated from the same data that were used for selecting the model. Sometimes, these biases may be substantial (see, for example, References 9.5 and 9.6). In the next section, we will show how one can examine whether the estimated regression coefficients and error mean square are biased to a substantial extent.

9.6 Model Validation

The final step in the model-building process is the validation of the selected regression models. Model validation usually involves checking a candidate model against independent data. Three basic ways of validating a regression model are:

1. Collection of new data to check the model and its predictive ability.
2. Comparison of results with theoretical expectations, earlier empirical results, and simulation results.
3. Use of a holdout sample to check the model and its predictive ability.

When a regression model is used in a controlled experiment, a repetition of the experiment and its analysis serves to validate the findings in the initial study if similar results for the regression coefficients, predictive ability, and the like are obtained. Similarly, findings in confirmatory observational studies are validated by a repetition of the study with other data.

As we noted in Section 9.1, there are generally no extensive problems in the selection of predictor variables in controlled experiments and confirmatory observational studies. In contrast, explanatory observational studies frequently involve large pools of explanatory variables and the selection of a subset of these for the final regression model. For these studies, validation of the regression model involves also the appropriateness of the variables selected, as well as the magnitudes of the regression coefficients, the predictive ability of the model, and the like. Our discussion of validation will focus primarily on issues that arise in validating regression models for exploratory observational studies. A good discussion of the need for replicating any study to establish the generalizability of the findings may be found in Reference 9.8. References 9.9 and 9.10 provide helpful presentations of issues arising in the validation of regression models.

Collection of New Data to Check Model

The best means of model validation is through the collection of new data. The purpose of collecting new data is to be able to examine whether the regression model developed from the earlier data is still applicable for the new data. If so, one has assurance about the applicability of the model to data beyond those on which the model is based.

Methods of Checking Validity. There are a variety of methods of examining the validity of the regression model against the new data. One validation method is to reestimate the model form chosen earlier using the new data. The estimated regression coefficients and various characteristics of the fitted model are then compared for consistency to those of the regression model based on the earlier data. If the results are consistent, they provide strong support that the chosen regression model is applicable under broader circumstances than those related to the original data.

A second validation method is designed to calibrate the predictive capability of the selected regression model. When a regression model is developed from given data, it is inevitable that the selected model is chosen, at least in large part, because it fits well the data at hand. For a different set of random outcomes, one may likely have arrived at a different model in terms of the predictor variables selected and/or their functional forms and interaction terms present in the model. A result of this model development process is that the error mean square *MSE* will tend to understate the inherent variability in making future predictions from the selected model.

A means of measuring the actual predictive capability of the selected regression model is to use this model to predict each case in the new data set and then to calculate the mean of the squared prediction errors, to be denoted by *MSPR*, which stands for *mean squared prediction error:*

$$MSPR = \frac{\sum_{i=1}^{n^*}(Y_i - \hat{Y}_i)^2}{n^*} \tag{9.20}$$

where:

Y_i is the value of the response variable in the ith validation case

$\hat{Y}_i$ is the predicted value for the ith validation case based on the model-building data set

n^* is the number of cases in the validation data set

If the mean squared prediction error *MSPR* is fairly close to *MSE* based on the regression fit to the model-building data set, then the error mean square *MSE* for the selected regression model is not seriously biased and gives an appropriate indication of the predictive ability of the model. If the mean squared prediction error is much larger than *MSE*, one should rely on the mean squared prediction error as an indicator of how well the selected regression model will predict in the future.

Difficulties in Replicating a Study. Difficulties often arise when new data are collected to validate a regression model, especially with observational studies. Even with controlled experiments, however, there may be difficulties in replicating an earlier study in identical fashion. For instance, the laboratory equipment for the new study to be conducted in a different laboratory may differ from that used in the initial study, resulting in somewhat different calibrations for the response measurements.

The difficulties in replicating a study are particularly acute in the social sciences where controlled experiments often are not feasible. Repetition of an observational study usually involves different conditions, the differences being related to changes in setting and/or time. For instance, a study investigating the relation between amount of delegation of authority by executives in a firm to the age of the executive was repeated in another firm which has a somewhat different management philosophy. As another example, a study relating consumer purchases of a product to special promotional incentives was repeated in another year when the business climate differed substantially from that during the initial study.

It may be thought that an inability to reproduce a study identically makes the replication study useless for validation purposes. This is not the case. No single study is fully useful until we know how much the results of the study can be generalized. If a replication study for which the conditions of the setting differ only slightly from those of the initial study yields substantially different regression results, then we learn that the results of the initial study cannot be readily generalized. On the other hand, if the conditions differ substantially and the regression results are still similar, we find that the regression results can be generalized to apply under substantially varying conditions. Still another possibility is that the regression results for the replication study differ substantially from those of the initial study, the differences being related to changes in the setting. This information may be useful for enriching the regression model by including new explanatory variables that make the model more widely applicable.

Comment

When the new data are collected under controlled conditions in an experiment, it is desirable to include data points of major interest to check out the model predictions. If the model is to be used for making predictions over the entire range of the X observations, a possibility is to include data points that are uniformly distributed over the X space. ■

Comparison with Theory, Empirical Evidence, or Simulation Results

In some cases, theory, simulation results, or previous empirical results may be helpful in determining whether the selected model is reasonable. Comparisons of regression coefficients and predictions with theoretical expectations, previous empirical results, or simulation

results should be made. Unfortunately, there is often little theory that can be used to validate regression models.

Data Splitting

By far the preferred method to validate a regression model is through the collection of new data. Often, however, this is neither practical nor feasible. An alternative when the data set is large enough is to split the data into two sets. The first set, called the *model-building set* or the *training sample*, is used to develop the model. The second data set, called the *validation* or *prediction set*, is used to evaluate the reasonableness and predictive ability of the selected model. This validation procedure is often called *cross-validation*. Data splitting in effect is an attempt to simulate replication of the study.

The validation data set is used for validation in the same way as when new data are collected. The regression coefficients can be reestimated for the selected model and then compared for consistency with the coefficients obtained from the model-building data set. Also, predictions can be made for the data in the validation data set from the regression model developed from the model-building data set to calibrate the predictive ability of this regression model for the new data. When the calibration data set is large enough, one can also study how the "good" models considered in the model selection phase fare with the new data.

Data sets are often split equally into model-building and validation data sets. It is important, however, that the model-building data set be sufficiently large so that a reliable model can be developed. Recall in this connection that the number of cases should be at least 6 to 10 times the number of variables in the pool of predictor variables. Thus, when 10 variables are in the pool, the model-building data set should contain at least 60 to 100 cases. If the entire data set is not large enough under these circumstances for making an equal split, the validation data set will need to be smaller than the model-building data set.

Splits of the data can be made at random. Another possibility is to match cases in pairs and place one of each pair into one of the two split data sets. When data are collected sequentially in time, it is often useful to pick a point in time to divide the data. Generally, the earlier data are selected for the model-building set and the later data for the validation set. When seasonal or cyclical effects are present in the data (e.g., sales data), the split should be made at a point where the cycles are balanced.

Use of time or some other characteristic of the data to split the data set provides the opportunity to test the generalizability of the model since conditions may differ for the two data sets. Data in the validation set may have been created under different causal conditions than those of the model-building set. In some cases, data in the validation set may represent extrapolations with respect to the data in the model-building set (e.g., sales data collected over time may contain a strong trend component). Such differential conditions may lead to a lack of validity of the model based on the model-building data set and indicate a need to broaden the regression model so that it is applicable under a broader scope of conditions.

A possible drawback of data splitting is that the variances of the estimated regression coefficients developed from the model-building data set will usually be larger than those that would have been obtained from the fit to the entire data set. If the model-building data set is reasonably large, however, these variances generally will not be that much larger than those for the entire data set. In any case, once the model has been validated, it is customary practice to use the entire data set for estimating the final regression model.

Example

In the surgical unit example, three models were favored by the various model-selection criteria. The SBC_p and $PRESS_p$ criteria favored the four-predictor model:

$$Y_i' = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_8 X_{i8} + \varepsilon_i \qquad \text{Model 1} \qquad \textbf{(9.21)}$$

C_p was minimized by the five-predictor model:

$$Y_i' = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_5 X_{i5} + \beta_8 X_{i8} + \varepsilon_i \qquad \text{Model 2} \qquad \textbf{(9.22)}$$

while the $R^2_{a,p}$ and AIC_p criteria were optimized by the six-predictor model:

$$Y_i' = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_8 X_{i8} \qquad \text{Model 3} \qquad \textbf{(9.23)}$$

We wish to assess the validity of these three models, both internally and externally.

Some evidence of the internal validity of these fitted models can be obtained through an examination of the various model-selection criteria. Table 9.4 summarizes the fits of the three candidate models to the original (training) data set in columns (1), (3), and (5). We first consider the SSE_p, $PRESS_p$ and C_p criterion values. Recall that the $PRESS_p$ value is always larger than SSE_p because the regression fit for the ith case when this case is deleted in fitting can never be as good as that when the ith case is included. A $PRESS_p$

TABLE 9.4 Regression Results for Candidate Models (9.21), (9.22), and (9.23) Based on Model-Building and Validation Data Sets—Surgical Unit Example.

Statistic	(1) Model 1 Training Data Set	(2) Model 1 Validation Data Set	(3) Model 2 Training Data Set	(4) Model 2 Validation Data Set	(5) Model 3 Training Data Set	(6) Model 3 Validation Data Set
	5	5	6	6	7	7
	3.8524	3.6350	3.8671	3.6143	4.0540	3.4699
	0.1927	0.2894	0.1906	0.2907	0.2348	0.3468
	0.0733	0.0958	0.0712	0.0999	0.0715	0.0987
	0.0190	0.0319	0.0188	0.0323	0.0186	0.0325
	0.0142	0.0164	0.0139	0.0159	0.0138	0.0162
	0.0017	0.0023	0.0017	0.0024	0.0017	0.0024
	0.0155	0.0156	0.0151	0.0154	0.0151	0.0156
	0.0014	0.0020	0.0014	0.0020	0.0014	0.0021
	—	—	—	—	−0.0035	0.0025
	—	—	—	—	0.0026	0.0033
	—	—	0.0869	0.0731	0.0873	0.0727
	—	—	0.0582	0.0792	0.0577	0.0795
	0.3530	0.1860	0.3627	0.1886	0.3509	0.1931
	0.0772	0.0964	0.0765	0.0966	0.0764	0.0972
	2.1788	3.7951	2.0820	3.7288	2.0052	3.6822
	2.7378	4.5219	2.7827	4.6536	2.7723	4.8981
	5.7508	6.2094	5.5406	7.3331	5.7874	8.7166
	0.0445	0.0775	0.0434	0.0777	0.0427	0.0783
	0.0773	—	0.0764	—	0.0794	—
	0.8160	0.6824	0.8205	0.6815	0.8234	0.6787

TABLE 9.5 Potential Predictor Variables and Response Variable—Surgical Unit Example.

Case Number i	Blood-Clotting Score X_{i1}	Prognostic Index X_{i2}	Enzyme Test X_{i3}	Liver Test X_{i4}	Age X_{i5}	Gender X_{i6}	Alc. Use: Mod. X_{i7}	Alc. Use: Heavy X_{i8}	Survival Time Y_i	$Y_i' = \ln Y_i$
55	7.1	23	78	1.93	45	0	1	0	302	5.710
56	4.9	66	91	3.05	34	1	0	0	767	6.642
57	6.4	90	35	1.06	39	1	0	1	487	6.188
...	...	...	...	...	...	...	...	...	...	...
106	6.9	90	33	2.78	48	1	0	0	655	6.485
107	7.9	45	55	2.46	43	0	1	0	377	5.932
108	4.5	68	60	2.07	59	0	0	0	642	6.465

value reasonably close to SSE_p supports the validity of the fitted regression model and of MSE_p as an indicator of the predictive capability of this model. In this case, all three of the candidate models have $PRESS_p$ values that are reasonably close to SSE_p. For example, for Model 1, $PRESS_p = 2.7378$ and $SSE_p = 2.1788$. Recall also that if $C_p \approx p$, this suggests that there is little or no bias in the regression model. This is the case for the three models under consideration. The C_5, C_6, and C_7 values for the three models are, respectively, 5.7508, 5.5406, and 5.7874.

To validate the selected regression model externally, 54 additional cases had been held out for a validation data set. A portion of the data for these cases is shown in Table 9.5. The correlation matrix for these new data (not shown) is quite similar to the one in Figure 9.3 for the model-building data set. The estimated regression coefficients, their estimated standard deviations, and various model-selection criteria when regression models (9.21), (9.22), and (9.23) are fitted to the validation data set are shown in Table 9.4, columns 2, 4, and 6. Note the excellent agreement between the two sets of estimated regression coefficients, and the two sets of regression coefficient standard errors. For example, for Model 1 fit to the training data, $b_1 = .0733$; when fit to the validation data, we obtain $b_1 = .0958$. In view of the magnitude of the corresponding standard errors (.0190 and .0319), these values are reasonably close.

A review of Table 9.4 shows that most of the estimated coefficients agree quite closely. However, it is noteworthy that b_5 in Model 3—the coefficient of age—is negative for the training data ($b_5 = -0.0035$), and positive for the validation data ($b_5 = 0.0025$). This is certainly a cause for concern, and it raises doubts about the validity of Model 3.

To calibrate the predictive ability of the regression models fitted from the training data set, the mean squared prediction errors *MSPR* in (9.20) were calculated for the 54 cases in the validation data set in Table 9.5 for each of the three candidate models; they are .0773, .0764, and .0794, respectively. The mean squared prediction error generally will be larger than MSE_p based on the training data set because entirely new data are involved in the validation data set. In this case, the relevant MSE_p values for the three models are .0445, .0434, and .0427. The fact that *MSPR* here does not differ too greatly from MSE_p implies that the error mean square MSE_p based on the training data set is a reasonably valid indicator of the predictive ability of the fitted regression model. The closeness of the three *MSPR*

values suggest that the three candidate models perform comparably in terms of predictive accuracy.

As a consequence of the concerns noted earlier about Model 3, this model was eliminated from further consideration. The final selection was based on the principle of parsimony. While Models 1 and 2 performed comparably in the validation study, Model 1 achieves this level of performance with one fewer parameter. For this reason, Model 1 was ultimately chosen by the investigator as the final model.

Comments

1. Algorithms are available to split data so that the two data sets have similar statistical properties. The reader is referred to Reference 9.11 for a discussion of this and other issues associated with validation of regression models.

2. Refinements of data splitting have been proposed. With the *double cross-validation procedure,* for example, the model is built for each half of the split data and then tested on the other half of the data. Thus, two measures of consistency and predictive ability are obtained from the two fitted models. For smaller data sets, a procedure called *K-fold cross-validation* is often used. With this procedure, the data are first split into K roughly equal parts. For $k = 1, 2, \ldots, K$, we use the kth part as the validation set, fit the model using the other $k - 1$ parts, and obtain the predicted sum of squares for error. The K estimates of prediction error are then combined to produce a *K-fold cross-validation estimate*. Note that when $K = n$, the K-fold cross-validation estimate is the identical to the $PRESS_p$ statistic.

3. For small data sets where data splitting is impractical, the *PRESS* criterion in (9.17), considered earlier for use in subset selection, can be employed as a form of data splitting to assess the precision of model predictions. Recall that with this procedure, each data point is predicted from the least squares fitted regression function developed from the remaining $n - 1$ data points. A fairly close agreement between *PRESS* and *SSE* suggests that *MSE* may be a reasonably valid indicator of the selected model's predictive capability. Variations of *PRESS* for validation have also been proposed, whereby m cases are held out for validation and the remaining $n - m$ cases are used to fit the model. Reference 9.11 discusses these procedures, as well as issues dealing with optimal splitting of data sets.

4. When regression models built on observational data do not predict well outside the range of the X observations in the data set, the usual reason is the existence of multicollinearity among the X variables. Chapter 11 introduces possible solutions for this difficulty including ridge regression or other biased estimation techniques.

5. If a data set for an exploratory observational study is very large, it can be divided into three parts. The first part is used for model training, the second part for cross-validation and model selection, and the third part for testing and calibrating the final model (Reference 9.10). This approach avoids any bias resulting from estimating the regression parameters from the same data set used for developing the model. A disadvantage of this procedure is that the parameter estimates are derived from a smaller data set and hence are more imprecise than if the original data set were divided into two parts for model building and validation. Consequently, the division of a data set into three parts is used in practice only when the available data set is very large. ■

Cited References

9.1. Daniel, C., and F. S. Wood. *Fitting Equations to Data: Computer Analysis of Multifactor Data.* 2nd ed. New York: John Wiley & Sons, 1999.

9.2. Freedman, D. A. "A Note on Screening Regression Equations," *The American Statistician* 37 (1983), pp. 152–55.

9.3. Pope, P. T., and J. T. Webster. "The Use of an F-Statistic in Stepwise Regression," *Technometrics* 14 (1972), pp. 327–40.
9.4. Mantel, N. "Why Stepdown Procedures in Variable Selection," *Technometrics* 12 (1970), pp. 621–25.
9.5. Miller, A. J. *Subset Selection in Regression*. 2nd ed. London: Chapman and Hall, 2002.
9.6. Faraway, J. J. "On the Cost of Data Analysis," *Journal of Computational and Graphical Statistics* 1 (1992), pp. 213–29.
9.7. Breiman, L., and P. Spector. "Submodel Selection and Evaluation in Regression. The X-Random Case," *International Statistical Review* 60 (1992), pp. 291 -319.
9.8. Lindsay, R. M., and A. S. C. Ehrenberg. "The Design of Replicated Studies," *The American Statistician* 47 (1993), pp. 217–28.
9.9. Snee, R. D. "Validation of Regression Models: Methods and Examples," *Technometrics* 19 (1977), pp. 415–28.
9.10. Hastie, T., Tibshirani, R., and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2001.
9.11. Stone, M. "Cross-validatory Choice and Assessment of Statistical Prediction," *Journal of the Royal Statistical Society B* 36 (1974), pp. 111–47.

Problems

9.1. A speaker stated: "In well-designed experiments involving quantitative explanatory variables, a procedure for reducing the number of explanatory variables after the data are obtained is not necessary." Discuss.

9.2. The dean of a graduate school wishes to predict the grade point average in graduate work for recent applicants. List a dozen variables that might be useful explanatory variables here.

9.3. Two researchers, investigating factors affecting summer attendance at privately operated beaches on Lake Ontario, collected information on attendance and 11 explanatory variables for 42 beaches. Two summers were studied, of relatively hot and relatively cool weather, respectively. A "best" subsets algorithm now is to be used to reduce the number of explanatory variables for the final regression model.

a. Should the variables reduction be done for both summers combined, or should it be done separately for each summer? Explain the problems involved and how you might handle them.

b. Will the "best" subsets selection procedure choose those explanatory variables that are most important in a causal sense for determining beach attendance?

9.4. In forward stepwise regression, what advantage is there in using a relatively small α-to-enter value for adding variables? What advantage is there in using a larger α-to-enter value?

9.5. In forward stepwise regression, why should the α-to-enter value for adding variables never exceed the α-to-remove value for deleting variables?

9.6. Prepare a flowchart of each of the following selection methods: (1) forward stepwise regression, (2) forward selection, (3) backward elimination.

9.7. An engineer has stated: "Reduction of the number of explanatory variables should always be done using the objective forward stepwise regression procedure." Discuss.

9.8. An attendee at a regression modeling short course stated: "I rarely see validation of regression models mentioned in published papers, so it must really not be an important component of model building." Comment.

*9.9. Refer to **Patient satisfaction** Problem 6.15. The hospital administrator wishes to determine the best subset of predictor variables for predicting patient satisfaction.

a. Indicate which subset of predictor variables you would recommend as best for predicting patient satisfaction according to each of the following criteria: (1) $R^2_{a,p}$, (2) AIC_p, (3) C_p, (4) $PRESS_p$. Support your recommendations with appropriate graphs.

b. Do the four criteria in part (a) identify the same best subset? Does this always happen?

c. Would forward stepwise regression have any advantages here as a screening procedure over the all-possible-regressions procedure?

*9.10. **Job proficiency.** A personnel officer in a governmental agency administered four newly developed aptitude tests to each of 25 applicants for entry-level clerical positions in the agency. For purpose of the study, all 25 applicants were accepted for positions irrespective of their test scores. After a probationary period, each applicant was rated for proficiency on the job. The scores on the four tests (X_1, X_2, X_3, X_4) and the job proficiency score (Y) for the 25 employees were as follows:

Subject	Test Score				Job Proficiency Score
i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	Y_i
1	86	110	100	87	88
2	62	97	99	100	80
3	110	107	103	103	96
...	...	...	...	...	...
23	104	73	93	80	78
24	94	121	115	104	115
25	91	129	97	83	83

a. Prepare separate stem-and-leaf plots of the test scores for each of the four newly developed aptitude tests. Are there any noteworthy features in these plots? Comment.

b. Obtain the scatter plot matrix. Also obtain the correlation matrix of the X variables. What do the scatter plots suggest about the nature of the functional relationship between the response variable Y and each of the predictor variables? Are any serious multicollinearity problems evident? Explain.

c. Fit the multiple regression function containing all four predictor variables as first-order terms. Does it appear that all predictor variables should be retained?

*9.11. Refer to **Job proficiency** Problem 9.10.

a. Using only first-order terms for the predictor variables in the pool of potential X variables, find the four best subset regression models according to the $R^2_{a,p}$ criterion.

b. Since there is relatively little difference in $R^2_{a,p}$ for the four best subset models, what other criteria would you use to help in the selection of the best model? Discuss.

9.12. Refer to **Market share** data set in Appendix C.3 and Problem 8.42.

a. Using only first-order terms for predictor variables, find the three best subset regression models according to the SBC_p criterion.

b. Is your finding here in agreement with what you found in Problem 8.42 (b) and (c)?

9.13. **Lung pressure.** Increased arterial blood pressure in the lungs frequently leads to the development of heart failure in patients with chronic obstructive pulmonary disease (COPD). The standard method for determining arterial lung pressure is invasive, technically difficult, and involves some risk to the patient. Radionuclide imaging is a noninvasive, less risky method for estimating arterial pressure in the lungs. To investigate the predictive ability of this method, a cardiologist collected data on 19 mild-to-moderate COPD patients. The data that follow on the next page include the invasive measure of systolic pulmonary arterial pressure (Y) and three

potential noninvasive predictor variables. Two were obtained by using radionuclide imaging—emptying rate of blood into the pumping chamber of the heart (X_1) and ejection rate of blood pumped out of the heart into the lungs (X_2)—and the third predictor variable measures a blood gas (X_3).

a. Prepare separate dot plots for each of the three predictor variables. Are there any noteworthy features in these plots? Comment.

b. Obtain the scatter plot matrix. Also obtain the correlation matrix of the X variables. What do the scatter plots suggest about the nature of the functional relationship between Y and each of the predictor variables? Are any serious multicollinearity problems evident? Explain.

c. Fit the multiple regression function containing the three predictor variables as first-order terms. Does it appear that all predictor variables should be retained?

Subject i	X_{i1}	X_{i2}	X_{i3}	Y_i
1	45	36	45	49
2	30	28	40	55
3	11	16	42	85
...	...	...	...	...
17	27	51	44	29
18	37	32	54	40
19	34	40	36	31

Adapted from A. T. Marmor et al., "Improved Radionuclide Method for Assessment of Pulmonary Artery Pressure in COPD," *Chest* 89 (1986), pp. 64–69.

9.14. Refer to **Lung pressure** Problem 9.13.

a. Using first-order and second-order terms for each of the three predictor variables (centered around the mean) in the pool of potential X variables (including cross products of the first-order terms), find the three best hierarchical subset regression models according to the $R^2_{a,p}$ criterion.

b. Is there much difference in $R^2_{a,p}$ for the three best subset models?

9.15. **Kidney function.** Creatinine clearance (Y) is an important measure of kidney function, but is difficult to obtain in a clinical office setting because it requires 24-hour urine collection. To determine whether this measure can be predicted from some data that are easily available, a kidney specialist obtained the data that follow for 33 male subjects. The predictor variables are serum creatinine concentration (X_1), age (X_2), and weight (X_3).

Subject i	X_{i1}	X_{i2}	X_{i3}	Y_i
1	.71	38	71	132
2	1.48	78	69	53
3	2.21	69	85	50
...	...	...	..	
31	1.53	70	75	52
32	1.58	63	62	73
33	1.37	68	52	57

Adapted from W. J. Shih and S. Weisberg, "Assessing Influence in Multiple Linear Regression with Incomplete Data," *Technometrics* 28 (1986), pp. 231–40.

a. Prepare separate dot plots for each of the three predictor variables. Are there any noteworthy features in these plots? Comment.

b. Obtain the scatter plot matrix. Also obtain the correlation matrix of the X variables. What do the scatter plots suggest about the nature of the functional relationship between the response variable Y and each predictor variable? Discuss. Are any serious multicollinearity problems evident? Explain.

c. Fit the multiple regression function containing the three predictor variables as first-order terms. Does it appear that all predictor variables should be retained?

9.16. Refer to **Kidney function** Problem 9.15.

a. Using first-order and second-order terms for each of the three predictor variables (centered around the mean) in the pool of potential X variables (including cross products of the first-order terms), find the three best hierarchical subset regression models according to the C_p criterion.

b. Is there much difference in C_p for the three best subset models?

*9.17. Refer to **Patient satisfaction** Problems 6.15 and 9.9. The hospital administrator was interested to learn how the forward stepwise selection procedure and some of its variations would perform here.

a. Determine the subset of variables that is selected as best by the forward stepwise regression procedure, using F limits of 3.0 and 2.9 to add or delete a variable, respectively. Show your steps.

b. To what level of significance in any individual test is the F limit of 3.0 for adding a variable approximately equivalent here?

c. Determine the subset of variables that is selected as best by the forward selection procedure, using an F limit of 3.0 to add a variable. Show your steps.

d. Determine the subset of variables that is selected as best by the backward elimination procedure, using an F limit of 2.9 to delete a variable. Show your steps.

e. Compare the results of the three selection procedures. How consistent are these results? How do the results compare with those for all possible regressions in Problem 9.9?

*9.18. Refer to **Job proficiency** Problems 9.10 and 9.11.

a. Using forward stepwise regression, find the best subset of predictor variables to predict job proficiency. Use α limits of .05 and .10 for adding or deleting a variable, respectively.

b. How does the best subset according to forward stepwise regression compare with the best subset according to the $R^2_{a,p}$ criterion obtained in Problem 9.11a?

9.19. Refer to **Kidney function** Problems 9.15 and 9.16.

a. Using the same pool of potential X variables as in Problem 9.16a, find the best subset of variables according to forward stepwise regression with α limits of .10 and .15 to add or delete a variable, respectively.

b. How does the best subset according to forward stepwise regression compare with the best subset according to the $R^2_{a,p}$ criterion obtained in Problem 9.16a?

9.20. Refer to **Market share** data set in Appendix C.3 and Problems 8.42 and 9.12.

a. Using forward stepwise regression, find the best subset of predictor variables to predict market share of their product. Use α limits of .10 and .15 for adding or deleting a predictor, respectively.

b. How does the best subset according to forward stepwise regression compare with the best subset according to the SBC_p criterion used in 9.12a?

*9.21. Refer to **Job proficiency** Problems 9.10 and 9.18. To assess internally the predictive ability of the regression model identified in Problem 9.18, compute the *PRESS* statistic and compare it to *SSE*. What does this comparison suggest about the validity of *MSE* as an indicator of the predictive ability of the fitted model?

*9.22. Refer to **Job proficiency** Problems 9.10 and 9.18. To assess externally the validity of the regression model identified in Problem 9.18, 25 additional applicants for entry-level clerical positions in the agency were similarly tested and hired irrespective of their test scores. The data follow.

Subject	Test Score				Job Proficiency Score
i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	Y_i
26	65	109	88	84	58
27	85	90	104	98	92
28	93	73	91	82	71
..	.	...	...	..	.
48	115	119	102	94	95
49	129	70	94	95	81
50	136	104	106	104	109

a. Obtain the correlation matrix of the X variables for the validation data set and compare it with that obtained in Problem 9.10b for the model-building data set. Are the two correlation matrices reasonably similar?

b. Fit the regression model identified in Problem 9.18a to the validation data set. Compare the estimated regression coefficients and their estimated standard deviations to those obtained in Problem 9.18a. Also compare the error mean squares and coefficients of multiple determination. Do the estimates for the validation data set appear to be reasonably similar to those obtained for the model-building data set?

c. Calculate the mean squared prediction error in (9.20) and compare it to *MSE* obtained from the model-building data set. Is there evidence of a substantial bias problem in *MSE* here? Is this conclusion consistent with your finding in Problem 9.21? Discuss.

d. Combine the model-building data set in Problem 9.10 with the validation data set and fit the selected regression model to the combined data. Are the estimated standard deviations of the estimated regression coefficients appreciably reduced now from those obtained for the model-building data set?

9.23. Refer to **Lung pressure** Problems 9.13 and 9.14. The validity of the regression model identified as best in Problem 9.14a is to be assessed internally.

a. Calculate the *PRESS* statistic and compare it to *SSE*. What does this comparison suggest about the validity of *MSE* as an indicator of the predictive ability of the fitted model?

b. Case 8 alone accounts for approximately one-half of the entire *PRESS* statistic. Would you recommend modification of the model because of the strong impact of this case? What are some corrective action options that would lessen the effect of case 8? Discuss.

Exercise

9.24 The true quadratic regression function is $E\{Y\} = 15 + 20X + 3X^2$. The fitted linear regression function is $\hat{Y} = 13 + 40X$, for which $E\{b_0\} = 10$ and $E\{b_1\} = 45$. What are the bias and sampling error components of the mean squared error for $X_i = 10$ and for $X_i = 20$?