prediction equation, that is, how well will the equation predict on an independent sample(s) of data. The three methods of model validation, which are discussed in detail in section 3.11, are:

1. Data splitting—Randomly split the data, obtain a prediction equation on one half of the random split and then check its predictive power (cross-validate) on the other sample.
2. Use of the PRESS statistic.
3. Obtain an *estimate* of the average predictive power of the equation on many other samples from the same population, using a formula due to Stein (Herzberg, 1969).

3.9 TWO COMPUTER EXAMPLES

To illustrate the use of several of the aforementioned model selection methods, we consider two computer examples. The first example illustrates the SPSS REGRESSION program, and uses data from Morrison (1983) on 32 students enrolled in an MBA course. We predict instructor course evaluation from 5 predictors. The second example illustrates SAS REG on quality ratings of 46 research doctorate programs in psychology, where we are attempting to predict quality ratings from factors such as number of program graduates, percentage of graduates that received fellowships or grant support, etc. (Singer & Willett, 1988).

Example 5—SPSS Regression on Morrison MBA Data

The data for this problem are from Morrison (1983). The dependent variable is instructor course evaluation in an MBA course, with the five predictors being clarity, stimulation, knowledge, interest, and course evaluation. We illustrate two of the sequential procedures, stepwise and backward selection, using the SPSSX REGRESSION program. The control lines for running the analyses, along with the correlation matrix, are given in Table 3.3.

SPSSX REGRESSION has "*p* values," denoted by PIN and POUT, which govern whether a predictor will enter the equation and whether it will be deleted. The default values are PIN = .05 and POUT = .10. In other words, a predictor must be "significant" at the .05 level to enter, or must not be significant at the .10 level to be deleted.

First, we discuss the stepwise procedure results. Examination of the correlation matrix in Table 3.3 reveals that three of the predictors (CLARITY, STIMUL, and COUEVAL) are strongly related to INSTEVAL (simple correlations of .862, .739, and .738, respectively). Because clarity has the highest correlation, it will enter the equation first. Superficially, it might appear that STIMUL or COUEVAL would enter next; however, we must take into account how these predictors are correlated with CLARITY, and indeed both have fairly high correlations with CLARITY (.617 and .651 respectively). Thus, they will not account for as much unique variance on INSTEVAL, above and beyond that of

TABLE 3.3
SPSS Control Lines for Stepwise and Backward Selection Runs on the Morrison MBA Data and the Correlation Matrix

```
TITLE 'MULTIPLE REGRESSION-5 PREDICTORS'.
DATA LIST FREE/INSTEVAL CLARITY STIMUL KNOWLEDGE INTEREST
  COUEVAL.
BEGIN DATA.
 1 1 2 1 1 2   1 2 2 1 1 1   1 1 1 1 1 2   1 1 2 1 1 2
 2 1 3 2 2 2   2 2 4 1 1 2   2 3 3 1 1 2   2 3 4 1 2 3
 2 2 3 1 3 3   2 2 2 2 2 2   2 2 3 2 1 2   2 2 2 3 3 2
 2 2 2 1 1 2   2 2 4 2 2 2   2 3 3 1 1 3   2 3 4 1 1 2
 2 3 2 1 1 2   3 4 4 3 2 2   3 4 3 1 1 4   3 4 3 1 2 3
 3 4 3 2 2 3   3 3 4 2 3 3   3 3 4 2 3 3   3 4 3 1 1 2
 3 4 5 1 1 3   3 3 5 1 2 3   3 4 4 1 2 3   3 4 4 1 1 3
 3 3 3 2 1 3   3 3 5 1 1 2   4 5 5 2 3 4   4 4 5 2 3 4
END DATA.
① REGRESSION DESCRIPTIVES = DEFAULT/
    VARIABLES = INSTEVAL TO COUEVAL/
  LIST.
  ② STATISTICS = DEFAULTS TOL SELECTION/
    DEPENDENT = INSTEVAL/
  ③ METHOD = STEPWISE/
  ④ CASEWISE = ALL PRED RESID ZRESID LEVER COOK/
  ⑤ SCATTERPLOT(*RES,*PRE)/.
```

CORRELATION MATRIX

	INSTEVAL	CLARITY	STIMUL	KNOWLEDGE	INTEREST	COUEVAL
INSTEVAL	1.000	.862	.739	.282	.435	.738
CLARITY	.862	1.000	.617	.057	.200	.651
STIMUL	.739	.617	1.000	.078	.317	.523
KNOWLEDGE	.282	.057	.078	1.000	.583	.041
INTEREST	.435	.200	.317	.583	1.000	.448
COUEVAL	.738	.651	.523	.041	.448	1.000

① The DESCRIPTIVES = DEFAULT subcommand yields the means, standard deviations and the correlation matrix for the variables.

② The DEFAULTS part of the STATISTICS subcommand yields, among other things, the ANOVA table for each step, R, R^2, and adjusted R^2. The HISTORY part is needed to obtain a summary table, which is very helpful to have.

③ To obtain the backward selection procedure, we would simply put METHOD = BACKWARD/

④ This CASEWISE subcommand yields important regression diagnostics: ZRESID (standardized residuals—for identifying outliers on y), LEVER (hat elements—for identifying outliers on predictors), and COOK (Cook's distance—for identifying influential data points).

⑤ This SCATTERPLOT subcommand yields the plot of the residuals vs. the predicted values, which is very useful for determining whether any of the assumptions underlying the linear regression model may be violated.

CLARITY, as first appeared. On the other hand, INTEREST, which has a considerably lower correlation with INSTEVAL (.44), is only correlated .20 with CLARITY. Thus, the variance on INSTEVAL it accounts for is relatively independent of the variance CLARITY accounted for. And, as seen in Table 3.4, it is INTEREST that enters the regression equation second.

STIMUL is the third and final predictor to enter, because its p value (.0086) is less than the default value of .05. Finally, the other predictors (KNOWLEDGE and COUEVAL) don't enter because their p values (.0989 and .1288) are greater than .05.

Selected printout from the backward selection procedure appears in Table 3.5. First, all of the predictors are put into the equation. Then, the procedure determines which of the predictors makes the *least* contribution when entered last in the equation. That predictor is INTEREST, and since its p value is .9097, it is deleted from the equation. None of the other predictors can be further deleted because their p values are much less than .10.

Interestingly, note that two *different* sets of predictors emerge from the two sequential selection procedures. The stepwise procedure yields the set (CLARITY, INTEREST, and STIMUL), where the backward procedure yields (COUEVAL, KNOWLEDGE, STIMUL, and CLARITY). However, CLARITY and STIMUL are common to both sets. On the grounds of parsimony, we might prefer the set (CLARITY, INTEREST, and STIMUL), especially because the adjusted R^2's for the two sets are quite close (.84 and .87).

There are three other things that should be checked out before settling on this as our chosen model:

1. We need to determine if the assumptions of the linear regression model are tenable.
2. We need an estimate of the cross validity power of the equation.
3. We need to check for the existence of outliers and/or influential data points.

Figure 3.4 shows the plot of the residuals versus the predicted values from SPSSX. This plot shows essentially random variation of the points about the horizontal line of 0, indicating no violations of assumptions.

The issues of cross-validity power and outliers are considered later in this chapter, and are applied to this problem in section 3.15, after both topics have been covered.

Example 6—SAS REG on Doctoral Programs in Psychology

The data for this example come from a National Academy of Sciences report (1982) that, among other things, provided ratings on the quality of 46 research doctoral programs in psychology. The six variables used to predict quality are:

NFACULTY—number of faculty members in the program as of December 1980.