1989

Multiple regression: chapter 9

Makridakis, Spyros

Wiley, John & Sons

Forecasting Methods for Management

SPYROS MAKRIDAKIS
The European Institute of Business Administration (INSEAD)

STEVEN C. WHEELWRIGHT
Graduate School of Business Administration, Harvard University

FIFTH EDITION

WILEY

JOHN WILEY & SONS
New York · Chichester · Brisbane · Toronto · Singapore

To our children

Aris and Petros

and Marianne, Michael, Melinda, Kristen, Matthew, and Spencer

Library of Congress Cataloging in Publication Data:

Makridakis, Spyros G.
Forecasting methods for management.

Includes index.
1. Economic forecasting. 2. Economic forecasting — methodology 3. Business forecasting. I. Wheelwright, Steven C. II. Title.
ISBN 0-471-60063-6

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CHAPTER 9

MULTIPLE REGRESSION

In Chapter 8 simple regression and correlation were introduced and discussed. In simple regression the basic proposition is that an independent variable can be used to predict the value of some dependent variable (the quantity to be forecast) on the basis of a linear relationship between the two variables. In the major example in that chapter the variable to be forecast was the number of orders received daily by a mail-order house. The independent variable on which that forecast was based was the weight of all mail for that day. In many decision-making situations more than one variable can be used to explain or forecast a certain dependent variable. For example, in the mail-order situation the day of the week, as well as the weight of mail received, might be used to predict the number of orders.

In situations where more than a single independent variable is necessary to forecast accurately, simple regression is not adequate. The idea of simple regression can be generalized, however, through the technique of multiple regression to allow the manager to include more than one independent variable. This chapter examines the extension and application of the basic principles of simple regression to situations in which several independent variables affect the outcome of some dependent variable.

The specific example we will use in this chapter to illustrate the principles and concepts of multiple regression and multiple correlation concerns the forecasting of annual sales for a company in the glass business. Table 9-1 lists some of the historical information that this company, California Plate Glass (CPG), has gathered.

This table contains data not only on the variable company sales (net sales), but also on two other variables, annual automobile production and the number of building contracts awarded annually. The management of CPG believes that its net sales are closely tied to these other two industries, since its major customers are automobile producers and building contractors. We assume that as a part of the planning process, top management has asked for a forecast of corporate sales on an annual basis for the next five years.

Table 9-1 Historical Data Relating to CPG Sales

Year	Net Sales, CPG, (Millions of Dollars)	Automobile Production (Millions of Units)	Building Contracts Awarded (Millions of Starts)
1972	280.0	3.909	9.43
1973	281.5	5.119	10.36
1974	337.4	6.666	14.50
1975	404.2	5.338	15.75
1976	402.1	4.321	16.78
1977	452.0	6.117	17.44
1978	431.7	5.559	19.77
1979	582.3	7.920	23.76
1980	596.6	5.816	31.61
1981	620.8	6.113	32.17
1982	513.6	4.258	35.09
1983	606.9	5.591	36.42
1984	629.0	6.675	36.58
1985	602.7	5.543	37.14
1986	656.7	6.933	41.30
1987	778.5	7.638	45.62
1988	877.6	7.752	47.38
1989 (est.)		6.400	48.51
1990 (est.)		7.900	51.23
1991 (est.)		8.400	57.47
1992 (est.)		8.600	61.03
1993 (est.)		8.900	66.25

Although the results of simple regression analysis may be satisfactory for forecasting sales, management probably would prefer to use the information it has on automobile production and building contracts at the same time; that is, since management knows that both factors are important and that they move somewhat independently of each other, it would like to be able to forecast net CPG sales as a function of both automobile production and building contracts awarded. Mathematically such a relationship could be written as

$$\text{net sales CPG} = f(\text{automobile production, building contracts awarded}).$$

This equation states that net sales for the company depend on two independent variables—automobile production and building contracts awarded. Although several different forms of the equation could be written to show the relation between these variables, a straightforward one would be

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 \tag{9-1}$$

where $\hat{Y}$ = estimated value of CPG annual sales
X_1 = annual automobile production
X_2 = annual building contracts awarded.

From this equation it can be seen that if either X_1 or X_2 were elimiated, we would have the same situation that we handled with simple linear regression. Since we have more than one independent variable (X_1 and X_2), the regression is known as *multiple*. Note that in Equation (9-1) the dependent variable (the one we wish to forecast) is expressed as a linear function of the independent variables X_1 and X_2.

Just as we used the method of least squares in Chapter 8 to find the coefficients a and b, we can use the same idea here to estimate the best values for a, b_1, and b_2. In simple linear regression that method amounted to fitting a straight line to the data points in a manner that minimized the sum of the squared errors. We represented that graphically by letting one axis represent Y and the other X. In the case of two independent variables, X_1 and X_2, we need a three-dimensional graph. The situation, however, is completely analogous to two dimensions, but we now have three axes, Y, X_1, and X_2, and we are trying to fit a plane to the data points available. We do that by minimizing the sum of the squared deviations from the plane.

In general, we could have several independent variables and we could still apply the method of least squares to solve for the values of $a, b_1, b_2, \ldots, b_k$. Multiple regression allows us to determine the estimated values of these parameters using the principle of least squares.

When we move beyond the case of simple regression, the computations and mathematics become quite complicated, although they follow the same basic concepts of simple regression. Because of this complexity, we will not go into the details of the formulas required to estimate the values of multiple regression parameters. We will assume that the manager has at his or her disposal a computer program for multiple regression that can handle all of these calculations. The use of multiple regression is not recommended, unless some kind of computer is available.

APPLICATION OF MULTIPLE REGRESSION ANALYSIS

To achieve a better understanding of the concept of multiple regression, we can use the data given in Table 9-1 and apply the method of least squares to obtain values for a, b_1, and b_2 in Equation (9-1). In the first step we state just

what the problem is and how we want to go about solving it. We assume that the task is to forecast the company's sales for the next five years (1989–1993) and that these forecasts will be based in part on the estimated values of automobile production and building contract awards for those years. Since we have a number of historical observations in Table 9-1, we would like to determine values for a, b_1, and b_2 on the basis of these historical values and then use Equation (9-1) to forecast the future values of company sales.

Using this historical information and a multiple regression computer program, we obtain the following results: $a = 19.1$, $b_1 = 35.7$, and $b_2 = 10.9$. Thus our equation for forecasting company sales can be written as

$$\hat{Y} = 19.1 + 35.7X_1 + 10.9X_2 \tag{9-2}$$

This states that on the basis of our historical observations (years 1972–1988), the best linear equation is the one shown in Equation (9-2). Note that the historical values we used in developing the equation were in millions of dollars for net sales, in millions of units for automobile production and in millions of starts for building contracts awarded. It is important to remember that the actual values of the parameters depend on the units that we used in estimating them. Thus it would be incorrect to interpret Equation (9-2) to mean that automobile production is much more important than building contracts in determining company sales, simply because 35.7 is larger than 10.9. If we had used different units for expressing automobile production, our coefficient for X_1 could have been smaller than our coefficient for X_2.

The proper interpretation of the values in Equation (9-2) is that when both X_1 and X_2 are 0, company sales Y will have a value of \$19.1 million, and that when automobile production increases by one million units, company sales will increase by \$35.7 million (other things, i.e. building contracts awarded, being held constant). Thus the coefficients in our equation generally provide the manager with a rough idea of how changes in each of the independent variables influence the value of the dependent variable Y. In order to forecast sales for each of the next five years we need to substitute estimated values for X_1 and X_2 in Equation (9-2). For the year 1989, these values are, for example, 6.4 and 48.51, respectively. Thus our estimate of sales for 1989 would be

$$\begin{aligned}\hat{Y} &= 19.1 + 35.7(6.4) + 10.9(48.51) \\ &= 776.3(\$ \text{ millions}).\end{aligned}$$

Similarly, the computations for years 1990 through 1993 can be made by using the appropriate values for automobile production and building contracts awarded.

It should be noted that this approach for forecasting requires that we have estimates of the values of the independent variables (in this case, X_1 and X_2). Thus in formulating a multiple regression equation, the manager will want to consider for which independent variables good estimates of future values will be available. The two variables used in this case would seem reasonable, since for economic reasons the country would most likely prepare long-range forecasts of those variables to help in general economic planning. The manager must keep in mind that the accuracy of the forecast for annual sales depends *in large part* on the accuracy of the forecast for building contracts awarded and automobile production. When these independent variables are in error, there is clearly going to be a compounding effect in terms of the error in the annual corporate sales forecast.

A final point about this example is that the forecasts for years 1989 through 1993 were made without first checking the significance of the parameters or the appropriateness of the equation on which those forecasts were to be based. These questions will be taken up in a later section.

MULTIPLE CORRELATION AND THE COEFFICIENT OF DETERMINATION

It will be recalled that in simple regression we computed a statistic called the coefficient of determination, which was simply the ratio of the explained variation to the total variation. The same ratio can also be computed in multiple regression, where again it is the explained variation over the total variation. This coefficient of determination, denoted by R^2, can take on values from 0 to 1, the latter representing a situation in which all the variation in Y is explained. The actual formula for calculating the coefficient of determination in this case is exactly the same as that used for simple regression:

$$R^2 = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2}. \tag{9-3}$$

Returning to the example of annual sales of CPG, we compute the coefficient of determination, using Equation (9-3), as 0.976. This means that 97.6% of the variation in annual sales can be explained by the combined variation in automobile production and building contracts awarded.

Table 9-2 Simple Correlation Matrix

	CPG Sales	Automobile Production	Building Contracts
CPG sales	1.000	0.688	0.948
Automobile production	0.688	1.000	0.530
Building contracts	0.948	0.530	1.000

In multiple regression it is possible to compute the individual coefficient of correlation for *each of the pairs of variables*. Thus a simple correlation coefficient could be computed for company sales and annual automobile production. Another simple correlation coefficient could be computed for annual sales and building contracts awarded. Finally, a correlation coefficient could be computed for annual automobile production and annual building contracts awarded. These three different correlation coefficients are usually referred to as the simple correlations, since they involve only two variables. They are most often represented in a correlation matrix like that shown in Table 9-2.

The simple correlation matrix is of value to the manager using multiple regression, because it indicates how each pair of variables is correlated. Thus most computer programs that perform multiple regression analysis include the computation of the simple correlation matrix. (Later in this chapter some of the uses of the simple correlation matrix are described.)

TESTS OF SIGNIFICANCE

An important question that must be answered before the results of multiple regression analysis can be used in forecasting future values is that of statistical significance. The computation of the coefficients in the regression equation is based on the use of a sample of historical observations. Consequently the reliability of forecasts based on that regression equation will depend largely on this specific sample of observations that were used in its development. Thus, the question of significance is really: how reliable are forecasts that are based on a multiple regression analysis of a given sample of data?

Although there are many tests of significance, three major ones should be mentioned in connection with multiple regression. The same three tests were discussed in Chapter 8 for simple regression.

The first test of significance that the manager should be concerned with in using multiple regression is a test that indicates the overall significance in the regression equation. The test used for this is the F statistic. (This test was described in Chapter 8 in connection with the significance of simple regression.)

The value of the F statistic is the ratio of the *explained variance* to the *unexplained variance*. This can be written mathematically in two equivalent forms. One form is

$$F = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2/(k-1)}{\Sigma(Y_i - \hat{Y}_i)^2/(n-k)} \tag{9-4}$$

where n = number of observations (data points)
k = number of coefficients.

Alternatively, it can be written as

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \tag{9-5}$$

where R^2 is the coefficient of determination.

Although both forms of this equation give the same numerical value for the F statistic, Equation (9-5) is generally easier to use because the coefficient of determination R^2 usually will have been calculated. In the example of the CPG Company we have already computed the coefficient of determination as $R^2 = 0.976$. Because we used 17 observations in determining the values of our parameters a, b_1, and b_2 and because we have three coefficients in our regression equation, Equation (9-5) yields

$$F = \frac{0.976/(3-1)}{(1-0.976)/(17-3)} = \frac{0.976}{0.024}\left(\frac{14}{2}\right) = 284.9.$$

For the F statistic the appropriate decision rule concerning significance at the 95% confidence level is that 284.9 be greater than the corresponding value from the table of F values. Since this value is 3.74, which is much smaller than 284.9, we can conclude that the regression equation is significant.

The second test involves testing the significance of the individual coefficients in the regression equation. Essentially, the question is whether the value of each coefficient is significantly different from 0 or whether it occurred by chance. This test consists of calculating the standard error for each of the coefficients and then using that error to determine whether the value of the coefficient is significantly different from 0.

The actual computation of the amount of standard error in each coefficient is generally included in the computer program that performs multiple regres-

Table 9-3 Tests of Significance of CPG Company Regression Equation

Coefficient	Coefficient Value	Standard Error	t test $\left(\frac{\text{Coefficient Value}}{\text{Standard Error}}\right)$	Value from Table ($\alpha = 0.05$)	Is Coefficient Value Significant?
a	19.1	51.9	0.37	2.145	No
b_1	35.7	10.1	3.55	2.145	Yes
b_2	10.9	0.97	11.17	2.145	Yes

sion. In most cases these results are given in the form of the t test for each of these coefficients. This t test can be used directly to determine the significance of each coefficient.

The results of the t test computations for the CPG sales example are given in Table 9-3. As we can see, the t test is simply the value of the coefficient divided by the standard deviation of that coefficient. Thus, it indicates the number of standard deviations that the computed value is different from 0. Table 9-3 shows that for a, the constant term in the regression equation, the computed value of 19.1 is only 0.37 standard deviation from 0. For b_1 and b_2 the number of standard deviations from 0 is much greater, 3.55 and 11.17, respectively.

The rule for determining whether a coefficient is significantly different from 0 at the 95% confidence level is that the absolute value of the computed t test must be greater than the corresponding value from the table.

In Table 9-3 it can be seen that the constant term a is *not* significantly different from 0, but both coefficients b_1 and b_2 are significantly different from 0. The fact that the value of a, the constant term, is not significantly different from 0 means that, on the basis of statistics, the manager has no reason to assume that the value of 19.1 is any more likely than a value of 0.

The third test of significance that the manager may wish to undertake entails calculating the standard error of a forecast. This allows confidence intervals to be developed around forecasts based on the regression line. Generally a 95% confidence interval is used. In Chapter 8 we developed the equation used for computing the standard error of forecast for simple regression. This equation represented the standard deviation of the size and the distance that the independent variables are from their mean values. The standard error of forecast for multiple regression is analogous to this, but since two or more independent variables are involved, it is difficult to visualize it graphically.

Because of the complexity of computing the standard error of forecast, this measure is generally included in the computer programs for multiple regression analysis. Once the standard error of forecast has been obtained, the manager can use it to develop a confidence interval around any forecast. For

example, the manager could have a 95% confidence level (assuming that the past pattern will remain the same during the forecasting phase) that the actual value would lie within ± 2 standard errors from the forecast value.

For the CPG example, the standard error of forecast associated with the mean value of the independent variables is 40.8. Thus if we wanted to prepare a forecast using the mean values of automobile production and building contracts awarded, we could be 95% confident that the actual value would fall in an interval of roughly ± 81.6 units around the forecast value. (Note that 81.6 equals two times the standard error of forecast.) The exact value of this interval could be found using the appropriate formula. Finally, it should be noted that the value ± 81.6 is in terms of millions of dollars, since those are the units of Y.

With each basic test of significance performed, the user will gain a better understanding of the multiple regression equation and the level of reliability that can be placed on the forecasts developed from it. However, managers must also be aware that, like all statistical methods, regression is built on certain assumptions. When those assumptions are violated, the technique can become unreliable and even misleading when applied in practice.

ASSUMPTIONS INHERENT IN MULTIPLE REGRESSION ANALYSIS

Four basic assumptions are made each time multiple regression is used in practice. An understanding of these assumptions and of the conditions necessary to meet them is important if regression analysis is to be used wisely.

In this section we discuss briefly each of these assumptions, the means of recognizing possible violations, and the methods of correcting them. Much has been written about the technical aspects of these assumptions, but this is generally beyond the scope of this text. The reader desiring additional information on these four points is referred to the Selected References at the end of this chapter.

The first assumption in regression analysis is that a linear relationship exists. This assumption states that the dependent variable is linearly related to each of the independent variables. (Technically, the assumption of linearity refers to linearity in the coefficients.) As shown in Chapter 8, a number of nonlinear relationships can be transformed into linear ones. Thus this restriction is not nearly so binding in practice as it may appear on the surface.

When the assumption of linearity is not met, the usual way of achieving linearity is to transform the variables into new variables that do exhibit linear relationships with Y. As a practical step, the manager is usually well advised to graph the relationships between the dependent variable Y and each

independent variable X_i to determine whether the linearity assumption has been met. An individual graph for each pair of variables can help identify any nonlinearities.

The second basic assumption in regression analysis is that of constant variance of the regression errors. This is often referred to by the technical name *homoscedasticity*. The technical term for the lack of constant variance is *heteroscedasticity*. This assumption states that the forecasting errors must be constant over the entire range of observations. In other words, the residuals e_i of the regression remain constant over the entire range from beginning to end. Figure 9-1(*a*) describes the kind of pattern that exists when constant variance is present. Figure 9-1(*b*) describes a situation in which the residuals increase as the value of the independent variable increases, and thus the assumption of constant variance is not met. This type of nonconstant variance is found often in real forecasting situations. Figure 9-1(*c*) presents a different kind of nonconstancy in the variance. Thus to meet the assumption of constant variance, a pattern like that shown in Figure 9-1(*a*) must exist.

The third basic assumption in regression is that the residuals are independent (random) of one another. This means that each residual value is independent of the values coming before and after it. In technical terms, when this assumption is not met, it is said that serial correlation (or autocorrelation) exists among successive residual values. The means of identifying independence of the residuals include a graphical representation of those values, examining the sign (plus or minus) of the residuals, or computing the Durbin–Watson (D–W) statistic. Figure 9-1(*c*), for instance, represents not only the lack of constant variance, but also a pattern in the residuals in which their signs change from plus to minus and back to plus as the value of the independent variable increases. The (D–W) statistic, which can be used to test for the presence of autocorrelation, is beyond the scope of this book, although a value of this test between about 1.5 and 2.5 implies an absence of autocorrelation among the residuals. Again computer programs usually include the computed value of the D–W statistic and the corresponding ranges (from a table of D–W values) that indicate whether or not the residuals are independent (random).

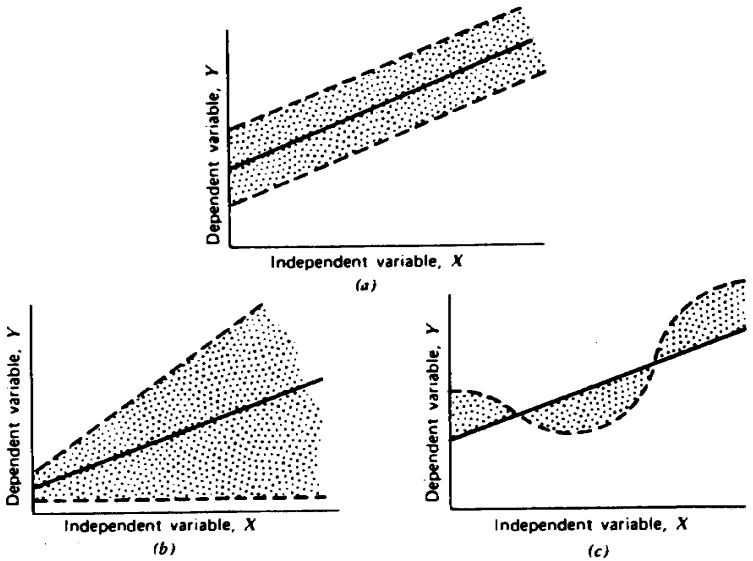

Figure 9-1 Constant Variance Assumption in Regression Analysis.

When the residuals are not independent, an important independent variable may have been omitted or a nonlinearity may exist among the variables used in the regression equation. Thus, rather than the equation capturing the basic underlying pattern with the residuals representing random errors, those residuals still include part of the basic pattern. If that pattern can be captured by the regression equation, more accurate forecasting is possible.

Two remedies are commonly used to eliminate autocorrelation in the residuals. First, an additional independent variable may be required to capture some of the variation in the dependent variable that could not be explained by existing independent variables and thus resulted in systematic, nonrandom errors. Second, the wrong functional form (such as, linear instead of exponential) may have been used in the regression equation. If neither a new variable nor a transformation of an existing variable can be devised that will eliminate autocorrelation, the method of first differences is often useful. Essentially, this method finds a new variable that has as its observed value the difference between each subsequent pair of observations for all variables. Thus if a series of observations with values 5, 8, 6, 4, and 7 were observed, the first differences for this set of data would be 3, -2, -2, and 3. If the first differences are computed for each of the variables in the regression equation, the regression coefficients can be recomputed using those differences as the observed values. (See Chapter 7, where the method of differencing was discussed as a way of eliminating trends from the data.)

When the independent variables exhibit strong autocorrelation, the variances are under- or overestimated, which makes the tests of significance invalid and the value of R^2 erroneous. The values of the regression coefficients $a, b_1, b_2, \ldots, b_k$ will be correct statistically and can be used, but nothing can be said about their significance as long as the residuals are autocorrelated.

The fourth basic assumption that the manager needs to consider in applying multiple regression analysis is that the residual values, if plotted, should be approximately normally distributed. This assumption is generally not restrictive, since the residuals represent the outcome of a large number of unimportant factors that influence the dependent variable, each to a relatively insignificant degree.

Thus, on the average, the influence of the factors will be cancelled out if the right model has been used. To check the assumption of normality one should plot the residuals and make sure they form a bell-shaped (normal) curve. If this assumption is not met, the tests of significance and the confidence intervals developed from them may be incorrect.

A final practical concern with multiple regression is the possibility of multicollinearity. Multicollinearity can develop when two or more of the independent variables are highly correlated. Technically the result is a near singular matrix, which has the same effect as trying to divide one number by another extremely small number. (You may recall that dividing a number by 0 gives a result of infinity.) If multicollinearity exists, the result is extremely large numbers that cannot be handled by the computer. The regression coefficient and all other output from the computer may, therefore, be erroneous. It should be stressed that multicollinearity is a computational (not a multiple regression) problem, because present-day computers are not big enough to handle the large numbers involved when two or more independent variables are highly correlated.

Multicollinearity is a real and frequent problem in economic and business data because of the high correlation between the different factors, such as population, GNP, personal disposable income, corporate sales, and corporate profits. One should be aware of its existence when selecting independent variables and when actually collecting data. The goal is to use independent variables that are not highly correlated (as a rule of thumb, the correlation between the independent variables included in the regression should not exceed $+0.7$ or be smaller than -0.7). If they are highly correlated, they provide redundent information that does not improve the explanatory power of the regression. [See Chapter 6 of Makridakis, Wheelwright and McGee (1989) for a more complete discussion of multicollinearity.]

USING MULTIPLE REGRESSION ANALYSIS IN PRACTICE

In the preceding sections of this chapter we have talked about the many considerations involved in applying multiple regression analysis and showed how the technique can be used in a straightforward example. In this section we bring together these different aspects relating to the application of regression analysis by developing a set of procedures that the manager can use and showing how these procedures can be used in a specific situation.

The fact that regression analysis is a forecasting technique based on understanding and measuring the extent of relationships means that the manager must identify those factors that appear to influence the variable to be forecast. One of the great advantages of multiple regression is that a number of different relationships can be hypothesized and tested with little effort when a computer program is available for doing so. Thus, the procedure that we outline will really go beyond the formulation of a regression equation and will describe how in a specific situation a manager might hypothesize certain relationships and then use regression analysis to determine which is the most appropriate. Nine basic steps are listed and described.

1. *Formulation of the Problem.* First the manager must state what the problem is and what it is that will be explained or predicted. This formulation should begin with a description of the decision-making situation and an identification of the variable or variables to be forecast rather than with the forecast itself. At the end of the formulation step a number of independent variables should have been identified and the dependent variable to be forecast should have been defined. This can be done by talking to people who are actually working in the area of concern and who are forecasting the dependent variable. Their experience and the factors they use when forecasting need to be considered when the manager is formulating the problem and hypothesizing possible solutions.

2. *Choice of Economic and Other Relevant Indicators.* Although problem formulation should identify some of the independent variables to be included, it is also necessary to identify additional possible influential factors and to determine which of them would be suitable for inclusion in the regression equation. This suitability must be based on the availability of data not only for historical periods but also for future periods for which the forecast is to be prepared. Some of the factors that are generally relevant include historical data relating to the company's operations and economic series relating to the general economy and the industry. Theoretically derived variables, as well as those identified through experience, must also be considered.

3. *Initial Test Run of Multiple Regression.* The initial run should include all the data on the independent and dependent variables and several transformations in case some of the relationships are not linear ones. It also may include the testing of a few plausible regression equations to observe the results that can be obtained. A useful output of this test run is the simple correlation matrix used in step 4.

4. *Studying the Matrix of Simple Correlations.* Careful selection of the variables, or their possible transformations, to include in the regression equation is fundamental to developing better forecasts from this method. The key is to pick independent variables (1) whose simple correlations are not bigger than 0.7 or smaller than -0.7 and (2) that add to the explanatory ability of the regression equation. The correlation coefficients between the independent and dependent variables being selected should be sufficiently larger than 0. (It should be remembered that the rule of thumb concerning multicollinearity does not always hold; this means that a value smaller than 0.7 or larger than -0.7 may result in multicollinearity, while a value larger than 0.7 or smaller than -0.7 may not.) At the end of this step the manager should have identified five or six alternative regression equations that seem promising and can be tested further.

5. *Deciding among Individual Regressions.* After a number of regression equations have been considered in step 4, a computer program should be used to estimate the coefficients of those regression equations on the basis of the data. For each of these regression equations, the manager can consider the significance of the entire regression, of the regression coefficients, and of the standard error of forecast. Once a regression equation has been found whose independent variables significantly influence the dependent variable, the usual procedure is to attempt to increase the R^2 value by introducing additional independent variables, checking each time to be sure that the tests of significance are still met.

6. *Observing the Value of R^2.* Once all regression coefficients have been found to be statistically significant and the standard error of the forecast is considered acceptable, the value of R^2 needs to be considered. R^2 tells us the percentage of variation in the dependent variable explained through the regression equation. If this percentage is small, the regression equation does not explain enough of the variation in the dependent variable. More independence may be required to explain the variation in the dependent variable and improve the value of R^2. The R^2 value provides a subjective measure that tells us the degree of the explanatory power of our regression equation. In some cases, as in medical research, unless R^2 is practically equal to 1, the regression equation cannot be used. In other cases even a small value of R^2 can be accepted as long as all regression coefficients are statistically significant and step 7 below has been completed.

7. *Checking the Validity of the Regression Assumptions.* Once a good equation (one that passes steps 5 and 6) has been identified, the manager must consider whether such a regression equation meets the four assumptions outlined in the preceding section. If it does not, appropriate steps should be taken to correct violations of the assumptions, or additional regression equations must be developed and tested. It must be remembered that high values of R^2 are meaningless when the D–W statistic is not in the appropriate range. If the test is not satisfied, it is advisable not to trust the regression equation no matter how large the value of R^2. Similarly, violation of the other assumptions can result in problems whose magnitude, however, is not as serious as when the D–W test indicates that the residuals are not random.

8. *Preparing a Forecast.* Once the manager has found a regression equation (1) whose regression coefficients are statistically significant, (2) that gives a sufficiently high value for R^2, and (3) that meets the assumptions inherent in regression, he or she can use the equation for forecasting purposes. In doing so, he or she should consider the confidence interval for individual forecasts and the accuracy of the values for the independent variable. As we pointed out earlier, most forecasts are based on estimated values of the independent variables rather than on actual values. Thus their validity needs to be determined, because if the forecasts of the independent variables are in error, the forecast of the dependent variable Y is also likely to be in error.

9. *Using the Regression Equation to Increase Understanding.* Quite often the biggest benefit from regression analysis is not in forecasting, but in explaining and helping us understand better some situations of interest. Consider, for instance, the regression equation we found for forecasting CPG's sales. This equation was

$$\hat{Y} = 19.1 + 35.7X_1 + 10.9X_2.$$

Suppose that the value of X_1 (automobile production) could not be forecast. Is the regression equation useless? Not at all. We know that for each additional million cars produced, the sales of CPG will increase by $35.7 million. This can be useful information by itself, because it can help us plan more effectively even though accurate forecasts of automobile production may not be available.

AN APPLICATION OF REGRESSION

As an example of how the preceding steps might be applied in developing an appropriate regression equation for forecasting, let us consider a company whose marketing manager wishes to forecast corporate sales for the coming year and to understand better the factors that influence them. The first step

Table 9-4 Forecasting Data for 1970 through 1988 (Semiannual)

(1) Personal Disposable Income (PDI) (Millions of Dollars)	(2) Dealers' Allowances (Thousands of Dollars)	(3) Price (Dollars)	(4) Product Development (Thousands of Dollars)	(5) Capital Investments (Thousands of Dollars)	(6) Advertising (Thousands of Dollars)	(7) Sales Expenses (Thousands of Dollars)	(8) Total Industry Advertising (Thousands of Dollars)	(9) Company Sales (Thousands of Dollars)
398	138	56.2058	12.1124	49.895	76.8621	228.80	98.205	5540.39
369	118	59.0443	9.3304	16.595	88.8056	177.45	224.953	5439.04
268	129	56.7236	28.7481	89.182	51.2972	166.40	263.032	4290.00
484	111	57.8627	12.8916	106.738	39.6473	258.05	320.928	5502.34
394	146	59.1178	13.3815	142.552	51.6517	209.30	406.989	4871.77
332	140	60.1113	11.0859	61.287	20.5476	180.05	246.996	4708.08
336	136	59.8398	24.9579	−30.385	40.1534	213.20	328.436	4627.81
383	104	60.0523	20.8096	−44.856	31.6456	200.85	298.456	4110.24
285	105	63.1415	8.4853	−28.373	12.4570	176.15	218.110	4122.69
277	135	62.3026	10.7301	75.723	68.3076	174.85	410.467	4842.25
456	128	64.9220	21.8473	144.030	52.4536	252.85	93.006	5740.65
355	131	64.8577	23.5062	112.904	76.6778	208.00	307.226	5094.10
364	120	63.5919	13.8940	128.347	96.0677	195.00	106.792	5383.20
320	147	65.6145	14.8659	10.097	47.9795	154.05	304.921	4888.17
311	143	67.0228	22.4940	−24.760	27.2319	180.70	59.612	4033.13
362	145	66.9049	23.3698	116.748	72.6681	219.70	238.986	4941.96
408	131	66.1843	13.0354	120.406	62.3129	234.65	141.074	5312.80
433	124	67.8651	8.0330	121.823	24.7122	258.05	290.832	5139.87
359	106	68.8892	27.0486	71.055	73.9126	196.30	413.636	4397.36
476	138	71.4177	18.2208	4.186	63.2737	278.85	206.454	5149.47
415	148	69.2775	7.7422	46.935	28.6762	207.35	79.566	5150.83
420	136	69.7334	10.1361	7.621	91.3635	213.20	428.982	4989.02
536	111	73.1628	27.3709	127.509	74.0169	296.40	273.072	5926.86
432	152	73.3650	15.5281	−49.574	16.1628	245.05	309.422	4703.88
436	123	73.0500	32.4918	100.098	42.9984	275.60	280.139	5365.59
415	119	74.9102	19.7127	−40.185	41.1346	211.25	314.548	4630.09
462	112	73.2007	14.8358	68.153	92.5180	282.75	212.058	5711.86
429	125	74.1615	11.3694	87.963	83.2870	217.75	118.065	5095.48
517	142	74.2838	26.7510	27.088	74.8921	306.80	344.553	6124.37
328	123	77.1409	19.6038	59.343	87.5103	210.60	140.872	4787.34
418	135	78.5910	34.6881	141.969	74.4712	269.75	82.855	5035.62
515	120	77.0938	23.2020	126.420	21.2711	328.25	398.425	5288.01
412	149	78.2313	35.7396	29.558	26.4941	258.05	124.027	4647.01
455	126	77.9296	21.5891	18.007	94.6311	232.70	117.911	5315.63
554	138	81.0394	19.5692	42.352	92.5448	323.70	161.250	6180.06
441	120	79.8485	15.5037	−21.558	50.0480	267.15	405.088	4800.97
417	120	80.6394	34.9238	148.450	83.1803	257.40	110.740	5512.13
461	132	82.2843	26.5496	−17.584	91.2214	266.50	170.392	5272.21

is to determine just why this forecast is needed and how it will be used. We will suppose that the marketing manager wants it for at least four reasons: (1) to supply her with estimates needed as her part in the corporate planning activity, (2) to give her an idea of the kind of staffing requirements she will have in sales and sales service to handle the company's increased sales, (3) to help in planning budget allocations for advertising, dealer discounts, and so on, and (4) to help her make better policy decisions concerning price, advertising, and product development expenditures.

With this initial identification of the problem, the marketing manager might well sit down with the sales manager and others in her marketing organization to determine the factors that might affect the company's sales. Let us suppose that they come up with the following model:

sales = f(personal disposable income, dealers' allowances, prices, product development expenditures, capital investments, advertising, sales expenses, total industry advertising, random effects).

Clearly, some of these factors will have a more important effect than others on the company's sales; others may turn out to be unimportant. Since any one of them, however, may have an important impact, it is useful to gather data on all of them at this early stage in the process. Thus the next step is to gather the information on these eight independent variables as well as on the dependent variable, company sales. Table 9-4 presents semiannual data covering the period from 1970 through 1988.

After these data have been collected, an initial multiple regression run can be made. As a starting point the regression equation

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \cdots + b_8X_8$$

can be used. This equation contains all eight independent variables. Some may not be important, but including them all initially gives a good starting basis. The results of applying a computer program by using the foregoing regression equation and the data in Table 9-4 are shown in Table 9-5. The second column in Table 9-5 lists the value of the constant term of the equation and the coefficient for each of the eight independent variables.

As can be seen in Table 9-5, not all the coefficients in this regression equation are significant. Looking at the t ratios in column 4 and comparing them to the corresponding values from the table, one can determine which coefficients are significant at the 95% level. It can be seen that independent

Table 9-5 Regression Equation for Semiannual Sales

(1) Variable	(2) Parameter Value	(3) Standard Error	(4) t Ratio	(5) Significant
Constant	2926.09	612.386	4.778	Yes
1 = PDI	3.809	1.528	2.491	Yes
2 = dealer allowances	5.064	3.138	1.613	No
3 = price	−17.126	7.998	−2.141	Yes
4 = product development expenditures	−10.258	6.274	−1.635	No
5 = capital investments	1.515	0.746	2.029	Yes
6 = advertising	8.053	1.778	4.528	Yes
7 = sales expenses	3.864	2.702	1.430	No
8 = total industry advertising	−0.539	0.377	−1.428	No

R^2 = 0.912; standard deviation of regression = 243.247;
Durbin–Watson statistic = 2.39146; F test = 1144.
t value from table (α = 0.05) = 1.96.
F value from table (α = 0.05) = 2.27

variables 2, 4, 7, and 8 (dealer allowances, product development expenditures, sales expenses, and total industry advertising, respectively) are not significantly different from 0 in terms of their impact on sales. This result could be due either to a lack of a significant relationship between variables 2, 4, 7, and 8 and Y or it could be due to multicollinearity between some of the variables, since both R^2 and the F test are large.

At this point the marketing manager can examine the simple correlation matrix shown in Table 9-6 to see how each independent variable is related to the company's sales. The bottom line of this table shows these correlations between the company sales (dependent variable) and each of the independent variables. We can see that variables 2, 4, and possibly 8 have a relatively small correlation with company sales. Variable 7, however, whose coefficient was not significant in Table 9-5, seems to have a fairly high correlation with company sales. The problem here is that multicollinearity does exist between variables 1 and 7. (The coefficient of correlation is 0.903.) This means that we need to drop either 1 or 7 from our regression equation. If we examine the correlation between sales and variable 1 and sales and variable 7, we see that the correlation is higher with variable 1. Thus we choose to eliminate variable 7 from our regression equation.

The marketing manager can now test an additional regression equation in which variables 2, 4, 7, and 8 have been eliminated. The new equation to be

Table 9-6 Simple Correlation Matrix for Semiannual Sales

Variable		(1) PDI	(2) Dealer Allowances	(3) Price	(4) Product Development	(5) Capital Investments	(6) Advertising	(7) Sales Expenses	(8) Total Industry Advertising	(9) Company Sales
	Variable Name									
1		1.000	−0.069	0.555	0.160	0.131	0.199	0.903	−0.020	0.742
2		−0.069	1.000	0.028	0.005	0.149	−0.119	−0.051	−0.145	0.009
3		0.555	0.028	1.000	0.438	−0.063	0.252	0.630	−0.182	0.285
4		0.160	0.005	0.438	1.000	0.217	0.102	0.361	−0.128	0.031
5		0.131	−0.149	−0.063	0.217	1.000	0.277	0.228	−0.063	0.410
6		0.199	−0.119	0.252	0.102	0.277	1.000	0.132	−0.197	0.526
7		0.903	−0.051	0.630	0.361	0.228	0.132	1.000	−0.019	0.667
8		−0.020	−0.145	−0.182	−0.128	−0.063	−0.197	−0.019	1.000	−0.175
9		0.742	0.009	0.285	0.031	0.410	0.526	0.667	−0.175	1.000

tested can be written as

$$\hat{Y} = a + b_1X_1 + b_3X_3 + b_5X_5 + b_6X_6.$$

The results for this regression analysis are presented in Table 9-7. As can be seen from this new regression equation, all the t ratios are statistically significant, which indicates that the term and each of the regression coefficients (b_1, b_3, b_5 and b_6) is significantly different from 0. If the F test is greater than 2.69, the entire regression equation is significant. Finally, the value of R^2 is equal to 0.781, which indicates that 78.1% of the fluctuations in sales are explained by the regression equation shown in Table 9-7. In practical terms the accuracy of this equation can be seen from Table 9-8, which gives the residuals (the difference between actual values and those predicted by the equation) and expresses those residuals as a percentage of the actual values. Since the greatest error is 14.76%, it can be assumed that the regression equation is quite adequate in explaining past sales. Furthermore, the standard deviation of regression has a value of about 260. This means that we can be 95% confident that our actual value will lie within ±$520,000 (± 2 standard deviations, that is ± 260,000 of the forecast value in the area of the mean.

Now the regression equation determined in Table 9-7 can be checked to see whether it conforms to the four basic assumptions. The equation does represent a linear relationship, the tests of significance are satisfied, the R^2 is good, and it appears that the regression equation is a good representation of this situation. The residuals are about constant. They are neither larger nor smaller in the beginning or at the end, as can be seen from Figure 9-2, which plots the residuals over time. The value of the D–W test is 2.31, which is close to the allowable range of 1.72 to 2.28. Finally, the residuals are about normally distributed, as shown in Figure 9-3, which plots the residuals in the form of a histogram.

The regression equation can now be used to prepare a forecast. This requires that values of X_1, X_3, X_5, and X_6 be estimated and then substituted in the regression equation to compute an estimate for Y. The equation also can be used to understand better the relative impact of at least a handful of factors on company sales. The precise equation taken from Table 9-7 is

$$\hat{Y} = 3276.55 + 5.70X_1 - 15.18X_3 + 1.55X_5 + 7.57X_6$$

where X_1 = personal disposable income
X_3 = price per ton (in dollars)
X_5 = capital investments (in thousands of dollars)
X_6 = advertising (in thousands of dollars)
$\hat{Y}$ = semiannual sales (in thousands of dollars).

Table 9-7 Regression Equation for Semiannual Sales

Dependent Variable is 9 = Sales

Variable	Parameter Value	Standard Error	*t* Ratio	Significant
Constant	3276.55	393.68500	8.32276	Yes
1 = PDI	5.69570	0.74394	7.65613	Yes
3 = price	−15.1783	6.96706	−2.17859	Yes
5 = capital investments	1.55114	0.70495	2.20033	Yes
6 = advertising	7.57419	1.77507	4.26698	Yes

$R^2 = 0.781$; $R = 0.884$; standard deviation of regression = 259.829; degrees of freedom = 33;
Durbin–Watson statistic = 2.31183; F test = 471.2
t value from table = 2.03.
F value from table = 2.69.

Table 9-8 Regression Results for Semiannual Sales (Predicted, Residuals and % Errors)

Actual	Predicted	Residuals	Percentage Error
5540.39	5349.89	190.501	3.43841E-02
5439.04	5180.44	258.598	4.75449E-02
4290.00	4468.90	−178.895	−4.17005E-02
5502.34	5620.87	−118.530	−2.15417E-02
4871.77	5235.68	−363.912	−7.46982E-02
4708.08	4505.83	202.256	4.29594E-02
4627.81	4539.03	88.783	1.91847E-02
4110.24	4717.04	−606.794	−0.14763
4122.69	3991.78	130.914	3.17545E-02
4842.25	4543.44	298.814	6.17097E-02
5740.65	5509.08	231.570	4.03385E-02
5094.10	5069.99	24.105	4.73197E-02
5383.20	5311.28	71.915	1.33593E-02
4888.17	4482.32	405.854	8.30277E-02
4033.13	4198.47	−165.336	−4.09943E-02
4941.96	5054.38	−112.418	−2.27476E-02
5312.80	5254.56	58.234	1.09612E-02
5139.87	5088.84	51.020	9.92636E-03
4397.36	4945.73	−548.365	−0.124703
5149.47	5376.45	−226.977	−4.40777E-02
5150.83	4733.14	417.690	8.10918E-02
4989.02	5314.13	−325.107	−6.51645E-02
5926.86	5977.36	− 50.498	−8.52030E-03
4703.88	4669.05	34.821	7.40262E-03
5365.59	5132.04	233.550	4.35273E-02
4630.09	4752.48	−122.386	−2.64328E-02
5711.86	5603.36	108.498	1.89951E-02
5095.48	5363.63	−266.150	−5.22326E-02
6124.37	5702.99	421.383	6.88044E-02
4787.34	4728.74	58.606	1.22419E-02
5035.62	5248.74	−213.121	−4.23227E-02
5288.01	5396.88	−108.870	−2.05881E-02
4647.01	4682.28	− 35.263	−7.58850E-03
5315.63	5429.94	−114.307	−2.15039E-02
6180.06	5968.57	211.487	3.42208E-02
4800.97	4922.02	−121.043	−2.52122E-02
5512.13	5287.97	244.161	4.06669E-02
5272.21	5316.98	− 44.769	−8.49161E-03

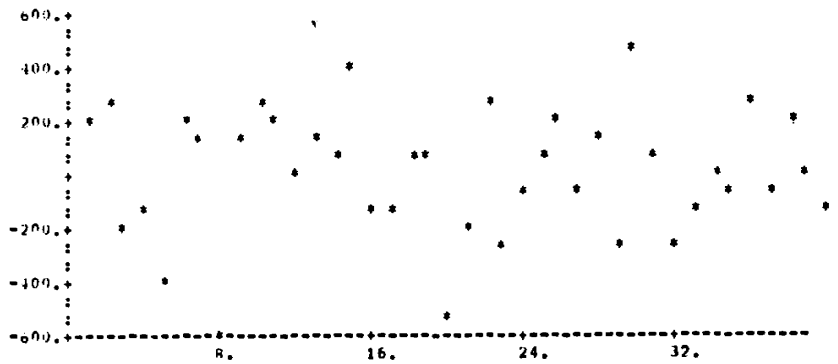

Figure 9-2 Plot of the Residuals for the Company Sales Regression Showing Their Constancy over Time.

The R^2 value of 0.781 tells us that the regression equation explains 78.1% of the total variation, that is, that variations in X_1, X_3, X_5, and X_6 explain 78.1% of the variation in the sales. The marketing manager must know the levels of precision and confidence that are associated with the values she inserts for the independent variables, to ensure that her forecast of semiannual sales will be as accurate as this regression equation allows. Thus the same types of approaches used to increase the accuracy of the sales forecast need to be applied to increase the accuracy of the estimates of the independent variables.

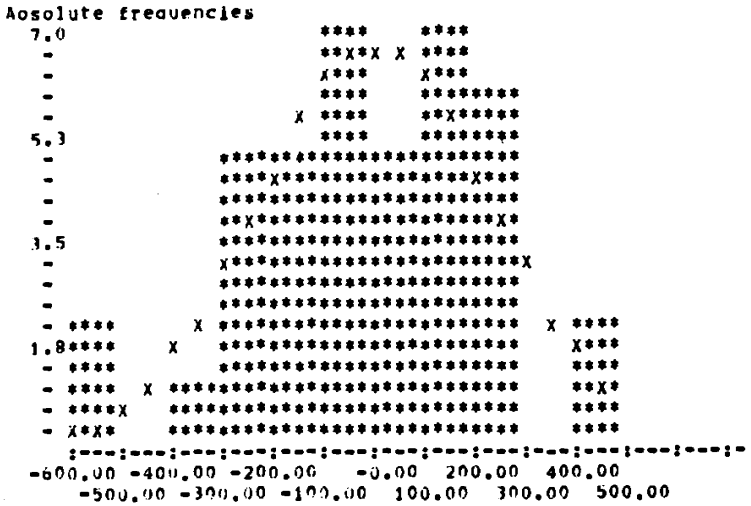

Figure 9-3 Histogram of the Residuals for the Company Sales Regression Showing that They Are Approximately Normally Distributed.

In addition, the marketing manager can gather a great deal of information from the regression equation. For instance, the constant value 3276.55 means that this portion of company sales is not explained by fluctuations in the four independent variables. If this constant term is large, it suggests that there may be additional independent variables that might explain more of the dependent variable. Identifying such variables also may increase the value of R^2, which indicates that 21.9% (100 − 78.1) of the total variation in sales is still unexplained.

Furthermore, the manager knows that as personal disposable income increases by $1 million (assuming that the remaining variables remain constant), company sales will increase by $5700. This is important information that might be used, for instance, for long-term planning purposes. Similarly, when capital investments and advertising increase, so do sales. It is interesting to note that for every dollar spent on advertising (assuming all other factors remain constant), the return is $1.55 in current period sales. This means that for every dollar spent on advertising, the return on sales is 55 cents more than what was spent, which is important information to consider in deciding how much to spend on advertising. Finally, it can be seen that the coefficient corresponding to price is negative. This means that when the price increases, sales will decrease. In fact, a $1.00 price increase decreases sales by $15.18 million. Similarly, a $1.00 price decrease will increase sales by $15.18 million. In spite of the caution that must be exercised in interpreting and using these results, regression analysis is an extremely powerful tool that can be used in a wide range of situations for both understanding and forecasting (see also Chapter 15).

SUMMARY

In the preceding sections we have considered some of the details of applying multiple regression in practice. There are also a number of general considerations that the manager should keep in mind in evaluating the appropriateness of this technique in comparison with other techniques. The major strength of multiple regression analysis is that it is an explanatory method that allows us to determine (estimate) virtually any kind of linear relationship that might exist between a dependent and one or more independent variables.

There are, of course, some drawbacks to the use of multiple regression.

One is that it requires estimates for the independent variables before a forecast can be made. Another is that most managers are reluctant to get into its details and to understand fully the power that it can bring to bear on a forecasting problem. (We hope this chapter will show that the method is very understandable and that by mastering some of the basic principles of its application, managers can use it wisely in a broad range of situations.)

Another potential drawback is the tendency to think that any time a high R^2 exists, the regression equation is automatically a good one. For this to be the case, the assumptions of regression must be satisfied and sufficient data must be available (at least 30 observations). A last point is that regression can be used reliably when and only when the relationship between the independent variables and the dependent variable does not change. If that relationship does change, it becomes necessary to collect a new set of data in order to redetermine the regression equation.

Given the substantial experience that has been gained by researchers and practitioners alike in the application of multiple regression, it is not surprising that a number of variations and modifications have been developed. Such things as stepwise regression (automatically selecting and then evaluating additional variables to be added to the basic regression equation), lead and lagged variables (shifting the time reference to create new variables), and dummy variables (creating variables with a value of 0 or 1, for example, to represent a seasonal factor) are just a few of these. The interested reader can pursue these in Makridakis, Wheelwright, and McGee (1989) and in Selected References.

SELECTED REFERENCES FOR FURTHER STUDY

Chatterjee, S., and B. Price, 1977. *Regression Analysis by Example*, Wiley, New York.

Clearly, P. J., and H. Levenbach, 1982. *The Professional Forecaster: The Forecasting Process through Data Analysis*, Lifetime Learning Publications, Belmont, CA.

Draper, N. R., and H. Smith, 1981. *Applied Regression Analysis*, 2nd ed., Wiley, New York.

Fildes, R., 1985. "Quantitative Forecasting—The State of the Art: Econometric Models," *Journal of Operational Research Society*, 36, no. 7, pp. 549–580.

Granger, C. W. J., 1980. *Forecasting in Business and Economics*, Academic Press, New York.

Gujarati, D., 1978. *Basic Econometrics*, McGraw-Hill, New York.

Hanke, J. E., and A. G. Reitsch, 1981. *Business Forecasting*, Allyn & Bacon, Boston.

Intrilligator, M., 1978. *Econometric Models, Techniques and Applications*, Prentice-Hall, Englewood Cliffs, NJ.

Johnston, J., 1972. *Econometric Methods*, Prentice-Hall, Englewood Cliffs, NJ.

Makridakis, S., and S.C. Wheelwright, 1978. *Interactive Forecasting*, 2nd ed., Holden-Day, San Francisco.

Makridakis, S., and S. C. Wheelwright (Eds.), 1979. *Forecasting. TIMS Studies in the Management Sciences*, vol. 12, North-Holland, Amsterdam.

Makridakis, S., S. C. Wheelwright, and V. E. McGee, 1989. *Forecasting: Methods and Applications*. 3rd ed., Wiley, New York.

Pindyck, R. S., and D. L. Rubinfeld, 1976. *Econometric Models and Economic Forecasts*, McGraw-Hill, New York.

Theil, H., 1971. *Principles of Econometrics*, Wiley, New York.

Wallace, D. T., and J. L. Silver, 1987. *Econometrics: An Introduction*, Addison-Wesley, Reading, MA.

Whetherill, G. B., 1986. *Regression Analysis with Applications*, Chapman and Hall, London.

Wonnacott, H., and R. J. Wonnacott, 1986. *Regression: A Second Course on Statistics*, Krieger, Melbourne.