CHAPTER 12 Multiple Regression

CHAPTER OUTLINE

12.1 The Multiple Regression Model
Model Specification
Model Objectives
Model Development
Three-Dimensional Graphing
12.2 Estimation of Coefficients
Least Squares Procedure
12.3 Explanatory Power of a Multiple Regression Equation
12.4 Confidence Intervals and Hypothesis Tests for Individual Regression Coefficients
Confidence Intervals
Tests of Hypotheses
12.5 Tests on Regression Coefficients
Tests on All Coefficients
Test on a Subset of Regression Coefficients
Comparison of F and t Tests
12.6 Prediction
12.7 Transformations for Nonlinear Regression Models
Quadratic Transformations
Logarithmic Transformations
12.8 Dummy Variables for Regression Models
Differences in Slope
12.9 Multiple Regression Analysis Application Procedure
Model Specification
Multiple Regression
Effect of Dropping a Statistically Significant Variable
Analysis of Residuals
12.10 Case Study Projects
Mini Case Studies
Nutrition Based Mini-Case Studies
Automobile Fuel Case Study Project

Introduction

In Chapter 11 we developed simple regression as a procedure for obtaining a linear equation that predicts a dependent or endogenous variable as a function of a single independent or exogenous variable—for example, total number of items sold as a function of price. However, in many situations, several independent variables jointly influence a dependent variable. Multiple regression enables us to determine the simultaneous effect of several independent variables on a dependent variable using the least squares principle.

Many important applications of multiple regression occur in business and economics. These applications include the following:

1. The quantity of goods sold is a function of price, income, advertising, price of substitute goods, and other variables.
2. Capital investment occurs when a business person believes that a profit can be made. Thus, capital investment is a function of variables related to the potential for profit, including interest rate, gross domestic product, consumer expectations, disposable income, and technological level.
3. Salary is a function of experience, education, age, and job rank.
4. Large retail, hotel, and restaurant companies decide on locations for new outlets based on the anticipated sales revenue and/or profitability. Using data from previous successful and unsuccessful locations, analysts can build models that predict sales or profit for a potential new location.

Business and economic analysis has some unique characteristics compared to analysis in other disciplines. Natural scientists work in a laboratory, where many—but not all—variables can be controlled. In contrast, the economist's and manager's laboratory is the world, and conditions cannot be controlled. Thus, we need tools such as multiple regression to estimate the simultaneous effect of several variables. Multiple regression as a "lab tool" is very important for the work of managers and economists. In this chapter we will see many specific applications in discussion examples and problem exercises.

The methods for fitting multiple regression models are based on the same least squares principle presented in Chapter 11, and, thus, the insights gained there extend directly to multiple regression. However, there are complexities introduced because of the relationships between the various exogenous variables. These require additional insights that are developed in this chapter.

12.1 The Multiple Regression Model

Our objective here is to learn how to use multiple regression for creating and analyzing models. Thus, we learn how multiple regression works and some guidelines for interpretation. A good understanding provides the capability for solving a wide range of applied problems. This study of multiple regression methods parallels the study of simple regression. The first step in model development is model specification, which includes the selection of model variables and the model form. Next, we study the least squares process, followed by an analysis of variability to identify the effects of each predictor variable. Then we study estimation, confidence intervals, and hypothesis testing. Computer applications are used extensively to indicate how the theory is applied to realistic problems. Your study of this material will be aided if you relate the ideas in this chapter to those presented in Chapter 11.

Model Specification

We begin with an application that illustrates the important task of regression model specification. Model specification includes selection of the exogenous variables and the functional form of the model.

Example 12.1 Process Manufacturing (Regression Model Specification)

The production manager for Flexible Circuits, Inc., has asked for your assistance in studying a manufacturing process. Flexible circuits are produced from a continuous roll of flexible resin material with a thin film of copper-conducting material bonded to its surface. Copper is bonded to the resin by passing the resin through a copper-based solution. The thickness of the copper is critical for high-quality circuits. Copper thickness depends, in part, on the temperature of the copper solution, speed of the production line, density of the solution, and thickness of the flexible resin material. To control the thickness of the bonded copper, the production manager needs to know the effect of each of these variables. You have been asked for assistance in developing a multiple regression model.

Solution Model development begins with a careful analysis of the problem context. The first step for this example would be an extended discussion with product design and manufacturing engineers so that you understand the process being modeled in detail. In some cases, you would study existing literature related to the process. The process must be understood and agreed to by the engineers and analysts before a useful model can be developed using multiple regression analysis. In this example the dependent variable, Y, is the copper thickness. Independent variables include temperature of the copper solution, X_1; speed of the production line, X_2; density of the solution, X_3; and thickness of the flexible resin material, X_4. These variables were identified as potential predictors of copper thickness, Y, by engineers and scientists that understand the technology of the plating process. Based on the study of the process, the resulting model specification is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

In this linear model the β_js are constant linear coefficients of the independent variables X_j that indicate the conditional effect of each independent variable on the determination of the dependent variable, Y, in the population. Thus, the coefficients β_j are parameters in the linear regression model. A series of production runs would then be made to obtain measurements of various combinations of independent and dependent variables. (See the discussion of experimental design in Section 13.2.)

Example 12.2 Store Location (Model Specification)

The director of planning for a large retailer was dissatisfied with the company's new-store development experience. In the past 4 years 25% of new stores failed to obtain their projected sales within the 2-year trial period and were closed, with substantial economic losses. The director wanted to develop better criteria for choosing store locations and decided that the historical experience of successful and unsuccessful stores should be studied.

Solution Discussion with a consultant indicated that data from stores that met and that did not meet anticipated sales could be used to develop a multiple regression model. The consultant suggested that the second year's sales should be used as the dependent variable, Y. A regression model would be used to predict second-year sales as a function of several independent variables that define the area surrounding the store. Stores would be located only where the predicted sales exceeded a minimum level. The model would also indicate the effect of various independent variables on sales.

After considerable discussion with people in the company, the consultant recommended the following independent variables:

1. X_1 = size of store
2. X_2 = traffic volume on highway in front of store
3. X_3 = stand-alone store versus shopping mall location
4. X_4 = location of competing store within 1/4 mile
5. X_5 = per capita income of population within 5 miles
6. X_6 = total number of people within 5 miles
7. X_7 = per capita income of population within 10 miles
8. X_8 = total number of people within 10 miles

Multiple regression was used to obtain estimates of the coefficients of the sales-prediction model from data collected for all stores opened during the past 8 years. The data set included both those stores that were still operating and those that were closed. A model was developed that could be used to predict second-year sales. This estimated equation included coefficient estimators, b_j, for the model parameters, β_j. To apply the estimated equation

$$\hat{y}_i = b_0 + \sum_{j=1}^{8} b_j x_{ji}$$

measurements of the independent variables were collected for each proposed new store location and the predicted sales were computed for that location. A predicted sales level was used, along with the judgment of marketing analysts and a committee of successful store managers, as input to the store location decision process.

Model Objectives

The strategy for model specification is influenced by the model objectives. One objective is prediction of a dependent or outcome variable. Applications include predicting or forecasting sales, output, total consumption, total investment, and many other business and economic performance criteria. A second objective is estimating the marginal effect of each independent variable. Economists and managers need to know how changes of independent variables, X_j, where $j = 1, \dots, K$, change performance measures, Y. For example, consider the following:

1. How do sales change as a result of a price increase and advertising expenditures?
2. How does output change when the amounts of labor and capital are changed?
3. Does infant mortality become lower when health care expenditures and local sanitation are increased?

Regression Objectives

Multiple regression provides two important results:

1. An estimated linear equation that predicts the dependent variable, Y, as a function of K observed independent variables, X_j, where $j = 1, \dots, K$:

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_K x_{Ki}$$

where $i = 1, \dots, n$ observations. The predicted value, $\hat{y}_i$, depends on the effect of the independent variables individually and their effect in combination with the other independent variables. Thus, we are interested in the combined effect of a particular combination of predictor variables.

2. The marginal change in the dependent variable, Y, that is related to changes in the independent variables—estimated by the coefficients, b_j. In multiple regression these coefficients depend on what other variables are included in the model. The coefficient b_j estimates the change in Y, given a unit change in X_j, while controlling for the simultaneous effect of the other independent variables.

In some problems both results are equally important. However, usually one will predominate (e.g., prediction of store sales, Y, in the store location example).

Marginal change is more difficult to estimate because the independent variables are related not only to the dependent variables but also to each other. If two or more independent variables change in a direct linear relationship with each other, it is difficult to determine the individual effect of each independent variable on the dependent variable.

Consider in detail the model in Example 12.2. The coefficient of x_5 indicates the change in sales for each unit change in the per capita income of the population within 5 miles, whereas that of x_7 indicates the sales change for change in per capita income of the population within 10 miles. It is, of course, likely that the variables x_5 and x_7 are correlated. Thus, to the extent that these variables both change at the same time, it is difficult to determine the contribution of each variable to change in store sales revenue. This correlation between independent variables introduces a complexity to the model. It is important to understand that the model predicts store sales revenue using the particular combination of variables contained in the model. The effect of a predictor variable is the effect of that variable when combined with the other variables. Thus, in general, the coefficient of a variable does not provide an indication of that variable's effect under all conditions. These complexities are explored further as we develop the multiple regression model.

Model Development

When applying multiple regression, we construct a model to explain variability in the dependent variable. In order to do this, we want to include the simultaneous and individual influences of several independent variables. For example, suppose that we wanted to develop a model that would predict the annual profit margin for savings and loan associations using data collected over a period of years. An initial model specification indicated that the annual profit margin was related to the net revenue per deposit dollar and the number of savings and loan offices. The net annual revenue is expected to increase the annual profit margin, and the number of savings and loan offices is anticipated to decrease the annual profit margin because of increased competition. This would lead us to specify a population regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where

Y = annual profit margin
X_1 = net annual revenue per deposit dollar
X_2 = number of savings and loan offices for that year

Table 12.1 and the data file named **Savings and Loan** contain 25 observations by year of these variables. These data will be used to develop a linear model that predicts annual profit margin as a function of revenue per deposit dollar and number of offices (Spellman 1978).

But before we can estimate the model, we need to develop and understand the multiple regression procedure. To begin, let us consider the general multiple regression

Table 12.1 Savings and Loan Associations Operating Data

Year	Revenue per Dollar	Number of Offices	Profit Margin	Year	Revenue per Dollar	Number of Offices	Profit Margin
1	3.92	7,298	0.75	14	3.78	6,672	0.84
2	3.61	6,855	0.71	15	3.82	6,890	0.79
3	3.32	6,636	0.66	16	3.97	7,115	0.7
4	3.07	6,506	0.61	17	4.07	7,327	0.68
5	3.06	6,450	0.7	18	4.25	7,546	0.72
6	3.11	6,402	0.72	19	4.41	7,931	0.55
7	3.21	6,368	0.77	20	4.49	8,097	0.63
8	3.26	6,340	0.74	21	4.70	8,468	0.56
9	3.42	6,349	0.9	22	4.58	8,717	0.41
10	3.42	6,352	0.82	23	4.69	8,991	0.51
11	3.45	6,361	0.75	24	4.71	9,179	0.47
12	3.58	6,369	0.77	25	4.78	9,318	0.32
13	3.66	6,546	0.78				

model and note the differences from the simple regression model. The multiple regression model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

where ε_i is the random error term with a mean of 0 and a variance of σ^2, and the β_j terms are the coefficients, or marginal effects, of the independent, or exogenous variables, X_j, where $j = 1, \ldots, K$, given the effects of the other independent variables. The i terms indicate the observations with $i = 1, \ldots, n$. We use lowercase letters x_{ji} to denote specific values of variable X_j at observation i. We assume that the random errors ε_i are independent of the variables X_j and of each other to ensure proper estimates of the coefficients and their variances. In Chapter 13 we indicate the effect of relaxing these assumptions.

The sample estimated model is

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_K x_{Ki} + e_i$$

where e_i is the residual or difference between the observed value of Y and the estimated value of Y obtained by using the estimated coefficients, b_j, where $j = 1, \ldots, K$. The regression procedure obtains simultaneous estimates, b_j, of the population model coefficients, β_j, using the least squares procedure.

In our savings and loan associations example, the population model for individual data points is as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

This reduced model with only two predictor variables provides the opportunity for developing additional insights into the regression procedure. The regression function can be depicted graphically in three dimensions, as shown in Figure 12.1. The regression function is shown as a plane whose Y values are a function of the independent variable values of X_1 and X_2. For each possible pair, x_{1i}, x_{2i}, the expected value of the dependent variable, Y, is on the plane. Figure 12.2 specifically illustrates the savings and loan example. An increase in X_1 leads to an increase in the expected value of Y, conditional on the effect of X_2. Similarly, an increase in X_2 leads to a decrease in the expected value of Y, conditional on the effect of X_1.

To complete our model, we add an error term defined as ε. This error term recognizes that no postulated relationship will hold exactly and that there are likely to be additional variables that also affect the observed value of Y. Thus, in the application setting we observe

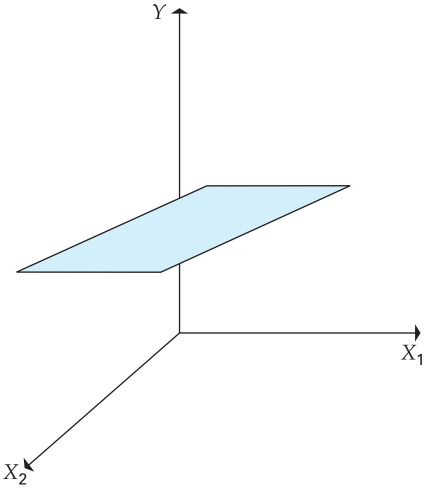


Figure 12.1 The Plane Is the Expected Value of Y as a Function of X_1 and X_2

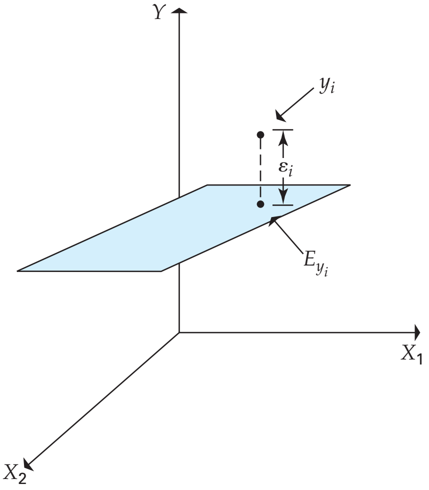


Figure 12.2 Comparison of the Observed and Expected Values of Y as a Function of Two Independent Variables

the expected value of the dependent variable, Y—as depicted by the plane in Figure 12.2—plus a random error term, ε, that represents the portion of Y not included in the expected value. As a result, the data model has the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

The Population Multiple Regression Model

The **population multiple regression model** defines the relationship between a dependent, or endogenous variable, Y, and a set of independent, or exogenous, variables, X_j, where $j = 1, \ldots, K$. The x_{ji} terms are assumed to be fixed numbers; Y is a random variable *with* y_i defined for each observation, i, where $i = 1, \ldots, n$ and n is the number of observations. The model is defined as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \qquad (12.1)$$

where the β_j terms are constant coefficients and the instances of ε_i are random variables with a mean of 0 and a variance of σ^2.

For the savings and loan example, with two independent variables, the population regression model is as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Given particular values of the net percentage revenue, x_{1i}, and the number of savings and loan offices, x_{2i}, the observed profit margin, y_i, is the sum of two parts: the expected value, $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$, and the random error term, ε_i. The random error term can be regarded as the combination of the effects of numerous other unidentified factors that affect profit margins. Figure 12.2 illustrates the model, with the plane indicating the expected value for various combinations of the independent variables and with the ε_i, shown as the deviation between the expected value, and the observed value of Y, marked by a large dot, for a particular data point. In general, the observed values of Y will not lie on the plane but instead will be above or below the plane because of the positive or negative error terms, ε_i.

Simple regression, developed in the previous chapter, is merely a special case of multiple regression with only one predictor variable, and, hence, the plane is reduced to a line. Thus, the theory and analysis developed for simple regression also apply to multiple regression. However, there are some additional interpretations that we will develop in our study of multiple regression. One of the important interpretations is illustrated in the following discussion of three-dimensional graphing.

Three-Dimensional Graphing

Your understanding of the multiple regression procedure might be helped by considering a simplified graphical image. Look at the corner of the room in which you are sitting. The lines formed by the two walls and the floor represent the axes for two independent variables, X_1 and X_2. The corner between the two walls is the axis for the dependent variable, Y. To estimate a regression line, we collect sets of points $(x_{1i}, x_{2i}, \text{and } y_i)$.

Now, picture these points plotted in your room using the wall and floor corners as the three axes. With these points hanging in your room, we find a plane in space that comes close to all of them. This plane is the geometric form of the least squares equation. With these points in space we now maneuver a plane up and down and rotate it in two directions; all these shifts are done simultaneously until we have a plane that is "close" to all the points. Recall that we did this with a straight line in two dimensions in Chapter 11 to obtain the equation

$$\hat{y} = b_0 + b_1 x$$

Then, we extend that idea to three dimensions to obtain the equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

This process is, of course, more complicated compared to simple regression. But real problems are complicated, and regression provides a way to better analyze the complexity of these problems. We want to know how Y changes with changes in X_1. However, these changes are, in turn, influenced by the way X_2 changes. And if X_1 and X_2 have a fixed relationship with each other, we cannot tell how much each variable contributes to changes in Y.

Geometric interpretations of multiple regression become increasingly complex as the number of independent variables increases. However, the analogy to simple regression is extremely useful. We estimate the coefficients by minimizing the sum of squared deviations in the Y dimension about a linear function of the independent variables. In simple regression the function is a straight line on a two-dimensional graph. With two independent variables the function is a plane in three-dimensional space. Beyond two independent variables we have various complex hyperplanes that are impossible to visualize.

Exercises

Basic Exercises

12.1 Given the estimated linear model

$$\hat{y} = 10 + 3x_1 + 2x_2 + 4x_3$$

a. Compute $\hat{y}$ when $x_1 = 20, x_2 = 11$, and $x_3 = 10$.
b. Compute $\hat{y}$ when $x_1 = 15, x_2 = 14$, and $x_3 = 20$.
c. Compute $\hat{y}$ when $x_1 = 35, x_2 = 19$, and $x_3 = 25$.
d. Compute $\hat{y}$ when $x_1 = 10, x_2 = 17$, and $x_3 = 30$.

12.2 Given the estimated linear model

$$\hat{y} = 10 + 5x_1 + 4x_2 + 2x_3$$

a. Compute $\hat{y}$ when $x_1 = 20, x_2 = 11$, and $x_3 = 10$.
b. Compute $\hat{y}$ when $x_1 = 15, x_2 = 14$, and $x_3 = 20$.
c. Compute $\hat{y}$ when $x_1 = 35, x_2 = 19$, and $x_3 = 25$.
d. Compute $\hat{y}$ when $x_1 = 10, x_2 = 17$, and $x_3 = 30$.

12.3 Given the estimated linear model

$$\hat{y} = 10 + 2x_1 + 12x_2 + 8x_3$$

a. Compute $\hat{y}$ when $x_1 = 20, x_2 = 11, x_3 = 10$.
b. Compute $\hat{y}$ when $x_1 = 15, x_2 = 24, x_3 = 20$.
c. Compute $\hat{y}$ when $x_1 = 20, x_2 = 19, x_3 = 25$.
d. Compute $\hat{y}$ when $x_1 = 10, x_2 = 9, x_3 = 30$.

12.4 Given the following estimated linear model

$$\hat{y} = 10 + 2x_1 + 12x_2 + 8x_3$$

a. What is the change in $\hat{y}$ when x_1 increases by 4?
b. What is the change in $\hat{y}$ when x_3 increases by 1?
c. What is the change in $\hat{y}$ when x_2 increases by 2?

12.5 Given the following estimated linear model

$$\hat{y} = 10 - 2x_1 - 14x_2 + 6x_3$$

a. What is the change in $\hat{y}$ when x_1 increases by 4?
b. What is the change in $\hat{y}$ when x_3 decreases by 1?
c. What is the change in $\hat{y}$ when x_2 decreases by 2?

Application Exercises

12.6 An aircraft company wanted to predict the number of worker-hours necessary to finish the design of a new plane. Relevant explanatory variables were thought to be the plane's top speed, its weight, and the number of parts it had in common with other models built by the company. A sample of 27 of the company's planes was taken, and the following model was estimated:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

where

y_i = design effort, in millions of worker-hours
x_{1i} = plane's top speed, in miles per hour
x_{2i} = plane's weight, in tons
x_{3i} = percentage number of parts in common with other models

The estimated regression coefficients were as follows:

$$b_0 = 2 \quad b_1 = 0.661 \quad b_2 = 0.065 \quad b_3 = -0.018$$

Interpret these estimates.

12.7 In a study of the influence of financial institutions on bond interest rates in Germany, quarterly data over a period of 12 years were analyzed. The postulated model was

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

where

y_i = change over the quarter in the bond interest rates
x_{1i} = change over the quarter in bond purchases by financial institutions
x_{2i} = change over the quarter in bond sales by financial institutions

The estimated regression coefficients were as follows:

$$b_1 = 0.057 \quad b_2 = -0.065$$

Interpret these estimates.

12.8 The following model was fitted to a sample of 30 families in order to explain household milk consumption:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

where

y_i = milk consumption, in quarts per week
x_{1i} = weekly income, in hundreds of dollars
x_{2i} = family size

The least squares estimates of the regression parameters were as follows:

$$b_0 = -0.025 \quad b_1 = 0.052 \quad b_2 = 1.14$$

a. Interpret the estimates b_1 and b_2.
b. Is it possible to provide a meaningful interpretation of the estimate b_0?

12.9 The following model was fitted to a sample of 25 students using data obtained at the end of their freshman year in college. The aim was to explain students' weight gains:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \varepsilon_i$$

where

y_i = weight gained, in pounds, during freshman year
x_{1i} = average number of meals eaten per week
x_{2i} = average number of hours of exercise per week
x_{3i} = average number of beers consumed per week

The least squares estimates of the regression parameters were as follows:

$$b_0 = 7.35 \quad b_1 = 0.653 \quad b_2 = -1.345 \quad b_3 = 0.613$$

a. Interpret the estimates b_1, b_2, and b_3.
b. Is it possible to provide a meaningful interpretation of the estimate b_0?

12.2 Estimation of Coefficients

Multiple regression coefficients are computed using estimators obtained by the least squares procedure. This least squares procedure is similar to that presented in Chapter 11 for simple regression. However, the estimators are complicated by the relationships between the independent X_j variables that occur simultaneously with the relationships between the independent and dependent variables. For example, if two independent variables increase or decrease linearly with each other—a positive or negative correlation—while at the same time there are increases or decreases in the dependent variable, we cannot identify the unique effect of each independent variable to the change in the dependent variable. As a result, we will find that the estimated regression coefficients are less reliable if there are high correlations between two or more independent variables. The estimates of coefficients and their variances are always obtained using a computer. However, we will spend considerable effort studying the algebra and computational forms in least squares regression. This effort will provide you with the background to understand the procedure and to determine how different data patterns influence the results. We begin with the standard assumptions for the multiple regression model.

Standard Multiple Regression Assumptions

The population multiple regression model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

and we assume that n sets of observations are available. The following standard assumptions are made for the model:

1. The x_{ji} terms are fixed numbers, or they are realizations of random variables, X_j, that are independent of the error terms, ε_i. In the latter case, inference is carried out conditionally on the observed values of the x_{ji}s.
2. The expected value of the random variable Y is a linear function of the independent X_j variables.
3. The error terms are normally distributed random variables with a mean of 0 and the same variance, σ^2. The latter is called homoscedasticity, or uniform variance.

$$E[\varepsilon_i] = 0 \quad \text{and} \quad E[\varepsilon_i^2] = \sigma^2 \quad \text{for } (i = 1, \ldots, n)$$

4. The random error terms, ε_i, are not correlated with one another, so that

$$E[\varepsilon_i \varepsilon_l] = 0 \quad \text{for all } i \neq l$$

5. It is not possible to find a set of nonzero numbers, $c_1, \ldots, c_K$, such that

$$c_1 x_{1i} + c_2 x_{2i} + \cdots + c_K x_{Ki} = 0$$

This is the property of no direct linear relationship between the X_j variables.

The first four assumptions are essentially the same as those made for simple regression. The error terms in assumption 3 are assumed to be normally distributed for statistical inference. But we will see that just as with simple regression, the central limit theorem allows us to relax that assumption if the sample size is large enough. Assumption 5 excludes certain cases in which there are linear relationships between the predictor variables. For example, suppose we are interested in explaining the variability in rates charged for shipping corn. One obvious explanatory variable would be the distance the corn is shipped. Distance could be measured in several different units, such as miles or kilometers. But it would not make sense to use both distance in miles and distance in kilometers as predictor variables. These two measures are linear functions of each other and would not satisfy assumption 5. In addition, it would be foolish to try to assess their separate effects. As we shall see, the equations that compute the coefficient estimates and the computer programs will not work if assumption 5 is violated. In most cases, proper model specification will avoid violating assumption 5.

Least Squares Procedure

The least squares procedure for multiple regression computes the estimated coefficients so as to minimize the sum of the residuals squared. Recall that the residual is defined as

$$e_i = y_i - \hat{y}_i$$

where y_i is the observed value of Y and $\hat{y}_i$ is the value of Y predicted from the regression. Formally, we minimize SSE:

$$\begin{aligned} SSE &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{1i} + \cdots + b_K x_{Ki}))^2 \end{aligned}$$

This minimization is the process of finding a plane that best represents a set of points in space, as we considered in our discussion of three-dimensional graphing. To carry out the process formally, we use partial derivatives to develop a set of simultaneous normal equations that are then solved to obtain the coefficient estimators. For those with a good understanding of differential calculus, the chapter appendix presents some of the details of the process. However, one can obtain great insights by realizing that we want a linear equation that best represents the observed data, and this is accomplished by minimizing the squared deviations about the estimated regression equation. Fortunately, for the applications studied in this book, the complex computations are always performed using a statistical computer package such as Minitab, SAS, or SPSS. Our objective here is to understand how to interpret the regression results and use them to solve problems. We will do this by examining some of the intermediate algebraic results to help understand the effects of various data patterns on the coefficient estimators.

Least Squares Estimation of the Sample Multiple Regression

We begin with a sample of n observations denoted as $x_{1i}, x_{2i}, \ldots, x_{Ki}, y_i$, where $i = 1, \ldots, n$, measured for a process whose population multiple regression model is as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

The least squares estimates of the coefficients $\beta_1, \beta_2, \ldots, \beta_K$, are the values $b_0, b_1, \ldots, b_K$ for which the sum of the squared errors

$$SSE = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_K x_{Ki})^2 \tag{12.2}$$

is a minimum.

The resulting equation

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_K x_{Ki} \tag{12.3}$$

is the sample multiple regression of Y on $X_1, X_2, \ldots, X_K$.

Let us consider again the regression model with only two predictor variables.

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$$

The coefficient estimators are computed using the following equations:

$$b_1 = \frac{s_y(r_{x_1 y} - r_{x_1 x_2} r_{x_2 y})}{s_{x_1}(1 - r_{x_1 x_2}^2)} \tag{12.4}$$

$$b_2 = \frac{s_y(r_{x_2y} - r_{x_1x_2}r_{x_1y})}{s_{x_2}(1 - r^2_{x_1x_2})} \tag{12.5}$$

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 \tag{12.6}$$

where

r_{x_1y} is the sample correlation between X_1 and Y
r_{x_2y} is the sample correlation between X_2 and Y
$r_{x_1x_2}$ is the sample correlation between X_1 and X_2
s_{x_1} is the sample standard deviation for X_1
s_{x_2} is the sample standard deviation for X_2
s_y is the sample standard deviation for Y

In the equations for the coefficient estimators, we see that the slope coefficient estimate, b_1, not only depends on the correlation between Y and X_1 but also is affected by the correlation between X_1 and X_2 and the correlation between X_2 and Y. If the correlation between X_1 and X_2 is equal to 0, then the coefficient estimators, b_1 and b_2, will be the same as the coefficient estimator for simple regression—we should note that this hardly ever happens in business and economic analysis. Conversely, if the correlation between the independent variables is equal to 1, the coefficient estimators will be undefined, but this will result only from poor model specification and will violate multiple regression assumption 5. If the independent variables are perfectly correlated, then they both experience simultaneous relative changes. We see that in that case it is not possible to tell which variable predicts the change in Y. In Example 12.3 we see the effect of the correlations between independent variables by considering the savings and loan association problem, whose data are shown in Table 12.1.

Example 12.3 Profit Margins of Savings and Loan Associations (Regression Coefficient Estimation)

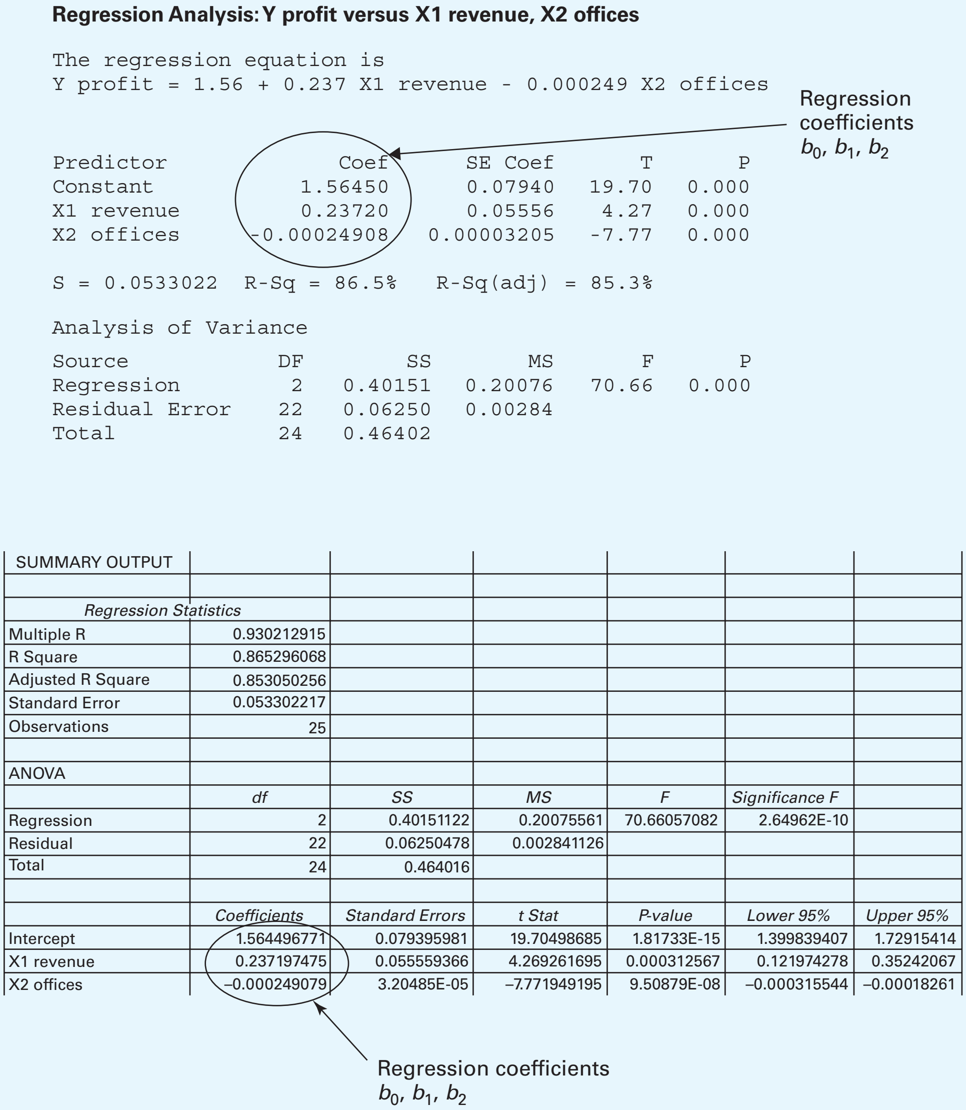
The director of the savings and loan association has asked you to compute the coefficients for variables that predict the percent profit margin.

Solution As a first step we develop a multiple regression model specification that predicts profit margin as a linear function of the net revenue per deposit dollar and the number of offices. Using the data in Table 12.1 that are stored in the **Savings and Loan** data file, we have estimated a multiple regression model, as seen in the Minitab and Excel outputs in Figure 12.3.

The estimated coefficients are identified in the computer output. We see that each unit increase in net revenue per deposit dollar, X_1, results in a 0.237 increase in profit margin—if the other variable does not change—and a unit increase in the number of offices decreases profit margin by 0.000249. Now consider the two simple regression models in Figures 12.4 and 12.5 with Y regressed on each independent variable by itself. First, consider Y regressed on revenue, X_1, in Figure 12.4. In this simple regression the coefficient for X_1 is -0.169, which is clearly different from $+0.237$ in multiple regression. We see that the correlation between X_1 and X_2 is 0.941. This large correlation has a major impact on the coefficient of X_1 in the multiple regression equation.

We see that the correlation between X_1 and X_2 is 0.941. Thus, the two variables tend to move together, and it is not surprising that the multiple regression coefficients are different from the simple regression coefficients.

Figure 12.3 Regression Equation for Savings and Loan Association Profit (Minitab and Excel Output)



Regression Analysis: Y profit versus X1 revenue, X2 offices

```
The regression equation is
Y profit = 1.56 + 0.237 X1 revenue - 0.000249 X2 offices

Predictor          Coef       SE Coef       T       P
Constant        1.56450       0.07940   19.70   0.000
X1 revenue      0.23720       0.05556    4.27   0.000
X2 offices  -0.00024908    0.00003205   -7.77   0.000

S = 0.0533022  R-Sq = 86.5%   R-Sq(adj) = 85.3%

Analysis of Variance

Source            DF        SS        MS       F       P
Regression         2   0.40151   0.20076   70.66   0.000
Residual Error    22   0.06250   0.00284
Total             24   0.46402
```

Regression coefficients b_0, b_1, b_2

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.930212915					
R Square	0.865296068					
Adjusted R Square	0.853050256					
Standard Error	0.053302217					
Observations	25					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	0.40151122	0.20075561	70.66057082	2.64962E-10	
Residual	22	0.06250478	0.002841126			
Total	24	0.464016				
	Coefficients	*Standard Errors*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.564496771	0.079395981	19.70498685	1.81733E-15	1.399839407	1.72915414
X1 revenue	0.237197475	0.055559366	4.269261695	0.000312567	0.121974278	0.35242067
X2 offices	–0.000249079	3.20485E-05	–7.771949195	9.50879E-08	–0.000315544	–0.00018261

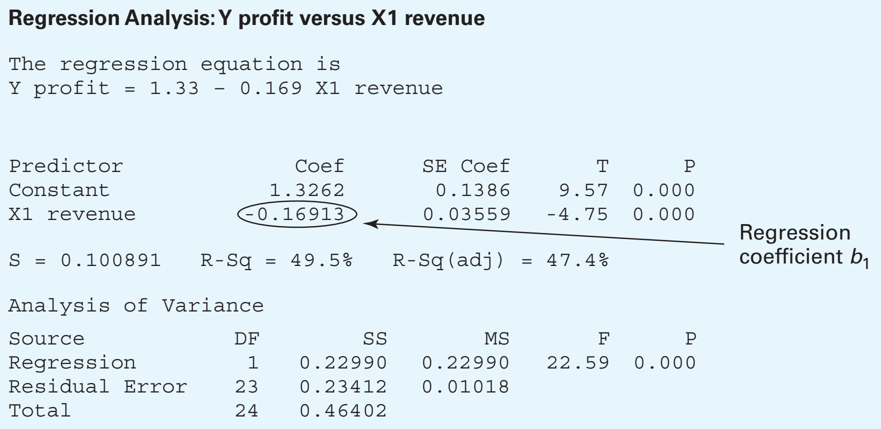
Regression coefficients b_0, b_1, b_2

Next, consider the regression of Y on X_2 alone in Figure 12.5. In this simple regression the slope coefficient for number of offices, X_2, is -0.000120, in contrast to -0.000249 for the multiple regression coefficient. This change in coefficients, while not quite as dramatic compared to the coefficient for X_1, also results from the high correlation between the independent variables.

The correlations between the three variables are as follows:

	Y Profit	X_1 Revenue
X_1 revenue	-0.704	
X_2 offices	-0.868	0.941

Figure 12.4 Savings and Loan Profit Regressed on Revenue

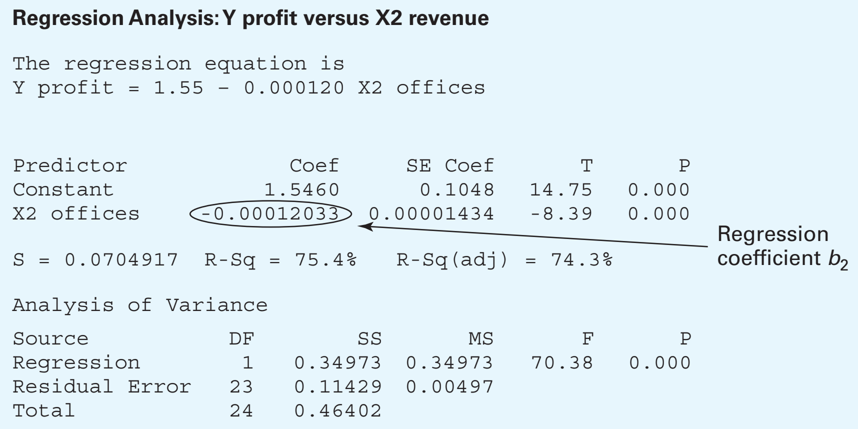


Regression Analysis: Y profit versus X1 revenue

```
The regression equation is
Y profit = 1.33 - 0.169 X1 revenue


Predictor           Coef     SE Coef      T      P
Constant          1.3262      0.1386   9.57  0.000
X1 revenue      -0.16913     0.03559  -4.75  0.000

S = 0.100891   R-Sq = 49.5%   R-Sq(adj) = 47.4%

Analysis of Variance

Source          DF       SS       MS      F      P
Regression       1  0.22990  0.22990  22.59  0.000
Residual Error  23  0.23412  0.01018
Total           24  0.46402
```

Regression coefficient b_1

Figure 12.5 Savings and Loan Profit Regressed on Number of Offices



Regression Analysis: Y profit versus X2 revenue

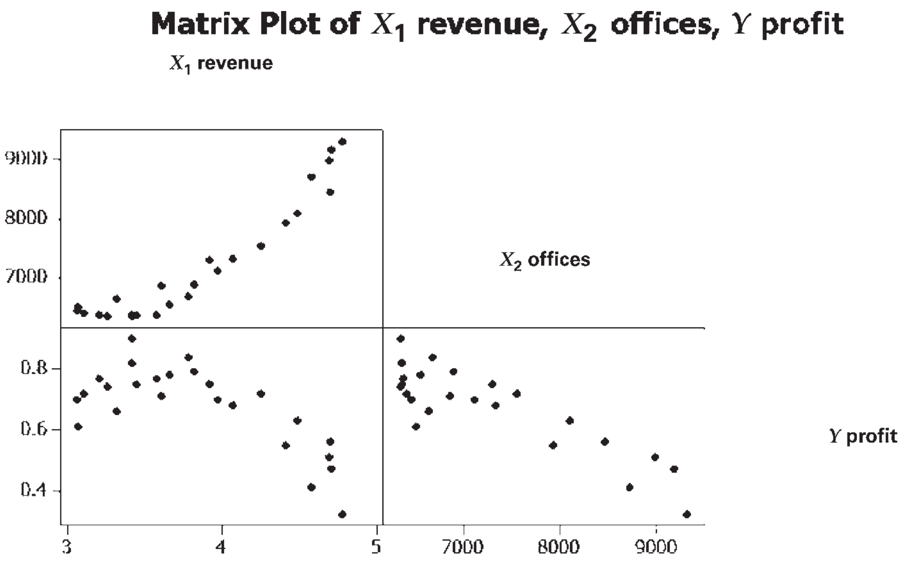
```
The regression equation is
Y profit = 1.55 - 0.000120 X2 offices


Predictor           Coef     SE Coef      T      P
Constant          1.5460      0.1048  14.75  0.000
X2 offices   -0.00012033  0.00001434  -8.39  0.000

S = 0.0704917  R-Sq = 75.4%   R-Sq(adj) = 74.3%

Analysis of Variance

Source          DF       SS       MS      F      P
Regression       1  0.34973  0.34973  70.38  0.000
Residual Error  23  0.11429  0.00497
Total           24  0.46402
```

Regression coefficient b_2

We should note that the multiple regression coefficients are *conditional coefficients*; that is, the estimated coefficient b_1 depends on the other independent variables included in the model. This will always be the case in multiple regression unless two independent variables have a sample correlation of zero—a very unlikely event.

These relationships can also be studied by using a "matrix plot" from Minitab, as shown in Figure 12.6. Matrix plots are not available in Excel. Note that the simple relationship between Y and X_2 is clearly linear, whereas the simple relationship between Y and X_1 is somewhat curvilinear. This nonlinear relationship between X_1 and Y explains in part why the coefficient of X_1 changed so dramatically from simple to multiple regression. We see from this example that correlations between independent variables can have a major influence on the estimated coefficients. Thus, if one has a choice, highly correlated independent variables should be avoided. But in many cases we do not have that choice. Regression coefficient estimates are always conditional on the other predictor variables in the model. In this example, profit margin increases as a function of net revenue per deposit dollar. However, the simultaneous increase in number of offices—which reduced profit—would hide the profit increase if a simple regression analysis were used. Thus, proper model specification—that is, choice of predictor variables—is very important. Model specification requires an understanding of the problem context and appropriate theory.

Figure 12.6 Matrix Plots for Savings and Loan Variables



Exercises

Visit **www.MyStatLab.com** or **www.pearsonhighered.com/newbold** to access the data files.

Basic Exercise

12.10 Compute the coefficients b_1 and b_2 for the regression model

$$\hat{y}_i = b_0 + b_1x_{1i} + b_2x_{2i}$$

given the following summary statistics.

a. $r_{x_1y} = 0.60, r_{x_2y} = 0.70, r_{x_1x_2} = 0.50,$ $s_{x_1} = 200, s_{x_2} = 100, s_y = 400$

b. $r_{x_1y} = -0.60, r_{x_2y} = 0.70, r_{x_1x_2} = -0.50,$ $s_{x_1} = 200, s_{x_2} = 100, s_y = 400$

c. $r_{x_1y} = 0.40, r_{x_2y} = 0.450, r_{x_1x_2} = 0.80,$ $s_{x_1} = 200, s_{x_2} = 100, s_y = 400$

d. $r_{x_1y} = 0.60, r_{x_2y} = -0.50, r_{x_1x_2} = -0.60,$ $s_{x_1} = 200, s_{x_2} = 100, s_y = 400$

Application Exercises

12.11 Consider the following estimated linear regression equations:

$$Y = a_0 + a_1X_1 \qquad Y = b_0 + b_1X_1 + b_2X_2$$

a. Show in detail the coefficient estimators for a_1 and b_1 when the correlation between X_1 and X_2 is equal to 0.

b. Show in detail the coefficient estimators for a_1 and b_1 when the correlation between X_1 and X_2 is equal to 1.

The following exercises require the use of a computer.

12.12 Amalgamated Power, Inc., has asked you to estimate a regression equation to determine the effect of various predictor variables on the demand for electricity sales. You will prepare a series of regression estimates and discuss the results using the quarterly data for electrical sales during the past 17 years in the data file **Power Demand**.

a. Estimate a regression equation with electricity sales as the dependent variable, using the number of customers and the price as predictor variables. Interpret the coefficients.

b. Estimate a regression equation (electricity sales) using only number of customers as a predictor variable. Interpret the coefficient and compare the result to the result from part a.

c. Estimate a regression equation (electricity sales) using the price and degree days as predictor variables. Interpret the coefficients. Compare the coefficient for price with that obtained in part a.

d. Estimate a regression equation (electricity sales) using disposable income and degree days as predictor variables. Interpret the coefficients.

12.13 Transportation Research, Inc., has asked you to prepare some multiple regression equations to estimate the effect of variables on fuel economy. The data for this study are contained in the data file **Motors**, and the dependent variable is miles per gallon—milpgal—as established by the Department of Transportation certification.

a. Prepare a regression equation that uses vehicle horsepower—horsepower—and vehicle weight—weight—as independent variables. Interpret the coefficients.

b. Prepare a second regression equation that adds the number of cylinders—cylinder—as an independent variable to the equation from part a. Interpret the coefficients.

c. Prepare a regression equation that uses number of cylinders and vehicle weight as independent variables. Interpret the coefficients and compare the results with those from parts a and b.

d. Prepare a regression equation that uses vehicle horsepower, vehicle weight, and price as predictor variables. Interpret the coefficients.

e. Write a short report that summarizes your results.

12.14 Transportation Research, Inc., has asked you to prepare some multiple regression equations to estimate the effect of variables on vehicle horsepower. The data for this study are contained in the data file **Motors**, and the dependent variable is vehicle horsepower—horsepower—as established by the Department of Transportation certification.

a. Prepare a regression equation that uses vehicle weight—weight—and cubic inches of cylinder displacement—displacement—as predictor variables. Interpret the coefficients.
b. Prepare a regression equation that uses vehicle weight, cylinder displacement, and number of cylinders—cylinder—as predictor variables. Interpret the coefficients and compare the results with those in part a.
c. Prepare a regression equation that uses vehicle weight, cylinder displacement, and miles per gallon—milpgal—as predictor variables. Interpret the coefficients and compare the results with those in part a.
d. Prepare a regression equation that uses vehicle weight, cylinder displacement, miles per gallon, and price as predictor variables. Interpret the coefficients and compare the results with those in part c.
e. Write a short report that presents the results of your analysis of this problem.

12.3 Explanatory Power of a Multiple Regression Equation

Multiple regression uses independent variables to explain the behavior of the dependent variable. We find that variability in the dependent variable can, in part, be explained by the linear function of the independent variables. In this section we develop a measure of the proportion of the variability in the dependent variable that can be explained by the multiple regression model.

The estimated regression model from the sample is

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_Kx_{Ki} + e_i$$

Alternatively, we can write

$$y_i = \hat{y}_i + e_i$$

where

$$\hat{y}_i = b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_Kx_{Ki}$$

is the predicted value of the dependent variable and the residual, e_i, is the difference between the observed and the predicted values. Table 12.2 contains these quantities for the savings and loan example in the first three columns.

We can subtract the sample mean of the dependent variable from both sides, giving

$$\begin{aligned}(y_i - \bar{y}) &= (\hat{y}_i - \bar{y}) + e_i\\ &= (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)\end{aligned}$$

which can be stated as follows:

observed deviation from mean = predicted deviation from mean + residual

Then by squaring both sides and summing over the index, i, we have

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y} + y_i - \hat{y}_i)^2\\ &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}e_i^2\end{aligned}$$

which is the sum-of-squares decomposition presented in Chapter 11:

$$SST = SSR + SSE$$

sum of squares total = sum of squares regression + sum of squares error

This simplified decomposition occurs because y_i and $\hat{y}_i$ are independent—y_i includes ε and $\hat{y}_i$ does not–and, thus,

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$$

Table 12.2 Actual Values, Predicted Values, and Residuals for Savings and Loan Regression

y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	$y_i - \overline{y}$	$\hat{y}_i - \overline{y}$
0.75	0.677	0.073	0.076	0.003
0.71	0.713	−0.003	0.036	0.039
0.66	0.699	−0.039	−0.014	0.025
0.61	0.672	−0.062	−0.064	−0.002
0.7	0.684	0.016	0.026	0.010
0.72	0.708	0.012	0.046	0.034
0.77	0.740	0.030	0.096	0.066
0.74	0.759	−0.019	0.066	0.085
0.9	0.794	0.106	0.226	0.120
0.82	0.794	0.026	0.146	0.120
0.75	0.798	−0.048	0.076	0.124
0.77	0.827	−0.057	0.096	0.153
0.78	0.802	−0.022	0.106	0.128
0.84	0.799	0.041	0.166	0.125
0.79	0.754	0.036	0.116	0.080
0.7	0.734	−0.034	0.026	0.060
0.68	0.705	−0.025	0.006	0.031
0.72	0.693	0.027	0.046	0.019
0.55	0.635	−0.085	−0.124	−0.039
0.63	0.613	0.017	−0.044	−0.061
0.56	0.570	−0.010	−0.114	−0.104
0.41	0.480	−0.070	−0.264	−0.194
0.51	0.437	0.073	−0.164	−0.237
0.47	0.395	0.075	−0.204	−0.279
0.32	0.377	−0.057	−0.354	−0.297
Sum of squares:		0.0625 (*SSE*)	0.4640 (*SST*)	0.4015 (*SSR*)

Sum-of-Squares Decomposition and the Coefficient of Determination

We begin with the multiple regression model fitted by least squares,

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_Kx_{Ki} + e_i = \hat{y}_i + e_i$$

where the b_j terms are the least squares estimates of the coefficients of the population regression model and the e terms are the residuals from the estimated regression model.

The model variability can be partitioned into the components

$$SST = SSR + SSE \tag{12.7}$$

where these components are defined as follows:

Sum-of-Squares Total

$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{12.8}$$

$$= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{12.9}$$

Sum-of-Squares Error

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}e_i^2 \tag{12.10}$$

Sum-of-Squares Regression or Explained Sum of Squares

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \quad (12.11)$$

This decomposition can be interpreted as follows:

total sample variability = explained variability + unexplained variability

The coefficient of determination, R^2, of the fitted regression is defined as the proportion of the total sample variability explained by the regression

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (12.12)$$

and it follows that

$$0 \leq R^2 \leq 1$$

The sum of squared errors is also used to compute the estimation for the variance of population model errors, as shown in Equation 12.13. As with simple regression, the variance of population errors is used for multiple regression statistical inference.

Estimation of Error Variance

Given the population multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

and the standard regression assumptions, let σ^2 denote the common variance of the error term, ε_i. Then an unbiased **estimate of error variance** is

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - K - 1} = \frac{SSE}{n - K - 1} \quad (12.13)$$

where K is the number of independent variables in the regression model. The square root of the variance, s_e, is also called the **standard error of the estimate**.

At this point we can also compute the mean square regression as follows:

$$MSR = \frac{SSR}{K}$$

We use MSR as a measure of the explained variability adjusted for the number of independent variables.

The sample mean for the savings and loan profit dependent variable is $\overline{y} = 0.674$, and we have used this value to compute the last two columns of Table 12.2. Using the data in Table 12.2 and the components, we can show that

$$SSE = 0.0625 \quad SST = 0.4640 \quad R^2 = 0.87$$

From these results we find that for this sample 87% of the variability in the savings and loan association's profit is explained by the linear relationships with net revenues and number of offices. Note that we could also compute the regression sum of squares from the identity

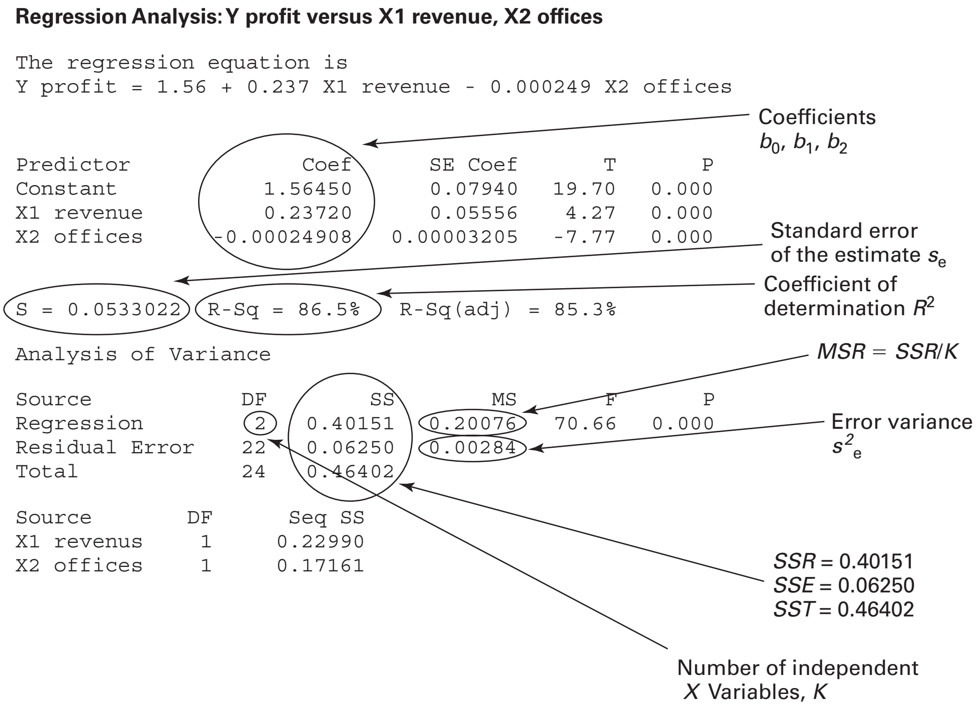
$$SSR = SST - SSE = 0.4640 - 0.0625 = 0.4015$$

We can also compute an estimate for the error variance σ^2 by using Equation 12.13:

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - K - 1} = \frac{SSE}{n - K - 1} = \frac{0.0625}{25 - 1 - 2} = 0.00284$$

Figure 12.7 presents the regression output from Minitab for the savings and loan association problem, with the various computed sums of squares indicated. These quantities are routinely computed by statistical computer packages, and the detail in Table 12.2 is included only to indicate how the sums of squares are computed. In all of the work that follows, we assume that the sums of squares are calculated by a computer package.

Figure 12.7 Regression Output for the Savings and Loan Association Problem



Regression Analysis: Y profit versus X1 revenue, X2 offices

```
The regression equation is
Y profit = 1.56 + 0.237 X1 revenue - 0.000249 X2 offices


Predictor              Coef      SE Coef      T      P
Constant            1.56450      0.07940  19.70  0.000
X1 revenue          0.23720      0.05556   4.27  0.000
X2 offices      -0.00024908   0.00003205  -7.77  0.000


S = 0.0533022  R-Sq = 86.5%   R-Sq(adj) = 85.3%

Analysis of Variance

Source           DF       SS       MS      F      P
Regression        2  0.40151  0.20076  70.66  0.000
Residual Error   22  0.06250  0.00284
Total            24  0.46402

Source        DF     Seq SS
X1 revenus     1    0.22990
X2 offices     1    0.17161
```

The components of variability have associated degrees of freedom. The SST quantity has $(n - 1)$ degrees of freedom because the mean of Y is required for its computation. The SSR component has K degrees of freedom because K coefficients are required for its computation. Finally, the SSE component has $(n - K - 1)$ degrees of freedom because K coefficients and the mean are required for its computation. Note that in Figure 12.7 the output includes the degrees of freedom (DF) associated with each component.

We routinely use the coefficient of determination, R^2, as a descriptive statistic to describe the strength of the linear relationship between the independent X variables and the dependent variable, Y. It is important to emphasize that R^2 can be used only to compare regression models that have the same set of sample observations of y_i, where $i = 1, \ldots, n$. This result is seen from the equation form as follows:

$$R^2 = 1 - \frac{SSE}{SST}$$

Thus, we see that R^2 can be large either because SSE is small—indicating that the observed points are close to the predicted points—or because SST is large. We have seen that SSE and s_e^2 indicate the closeness of the observed points to the predicted points. With the same SST for two or more regression equations, R^2 provides a comparable measure of the goodness of fit for the equations. This is the same result that was shown in the extended example in Section 11.4.

There is a potential problem with using R^2 as an overall measure of the quality of a fitted equation. As additional independent variables are added to a multiple regression model, the explained sum of squares, SSR, will increase—in essentially all applied situations— even if the additional independent variable is not an important predictor variable. Thus, we might find that R^2 has increased spuriously after one or more nonsignificant predictor variables have been added to the multiple regression model. In such a case, the increased value of R^2 would be misleading. To avoid this problem, the adjusted coefficient of determination can be computed as shown in Equation 12.14.

Adjusted Coefficient of Determination

The **adjusted coefficient of determination**, $\overline{R}^2$, is defined as follows:

$$\overline{R}^2 = 1 - \frac{SSE/(n-K-1)}{SST/(n-1)} \tag{12.14}$$

We use this measure to correct for the fact that nonrelevant independent variables will result in some small reduction in the error sum of squares. Thus, the adjusted $\overline{R}^2$ provides a better comparison between multiple regression models with different numbers of independent variables.

Returning to our savings and loan example, we see that

$$n = 25 \quad K = 2 \quad SSE = 0.0625 \quad SST = 0.4640$$

and, thus, the adjusted coefficient of determination is as follows:

$$\overline{R}^2 = 1 - \frac{0.0625/22}{0.4640/24} = 0.853$$

In this example the difference between R^2 and $\overline{R}^2$ is not very large. However, if the regression model had contained a number of independent variables that were not important conditional predictors, then the difference would be substantial. Another measure of relationship in multiple regression is the coefficient of multiple correlation.

Coefficient of Multiple Correlation

The **coefficient of multiple correlation** is the correlation between the predicted value and the observed value of the dependent variable

$$R = r(\hat{y}, y) = \sqrt{R^2} \tag{12.15}$$

and is equal to the square root of the multiple coefficient of determination. We use R as another measure of the strength of the relationship between the dependent variable and the independent variables. Thus, it is comparable to the correlation between Y and X in simple regression.

Exercises

Basic Exercises

12.15 A regression analysis has produced the following analysis of variance table:

Analysis of Variance			
Source	DF	SS	MS
Regression	3	4,500	
Residual error	26	500	

a. Compute s_e and s_e^2.
b. Compute SST.
c. Compute R^2 and the adjusted coefficient of determination.

12.16 A regression analysis has produced the following analysis of variance table:

Analysis of Variance			
Source	DF	SS	MS
Regression	2	7,000	
Residual error	29	2,500	

a. Compute s_e and s_e^2.
b. Compute SST.
c. Compute R^2 and the adjusted coefficient of determination.

12.17 A regression analysis has produced the following analysis of variance table:

Analysis of Variance			
Source	DF	SS	MS
Regression	4	40,000	
Residual error	45	10,000	

a. Compute s_e and s_e^2.
b. Compute SST.
c. Compute R^2 and the adjusted coefficient of determination.

12.18 A regression analysis has produced the following analysis of variance table:

Analysis of Variance			
Source	DF	SS	MS
Regression	5	80,000	
Residual error	200	15,000	

a. Compute s_e and s_e^2.
b. Compute SST.
c. Compute R^2 and the adjusted coefficient of determination.

Application Exercises

12.19 An aircraft company wanted to predict the number of worker-hours necessary to finish the design of a new plane. Relevant explanatory variables were thought to be the plane's top speed, its weight, and the number of parts it had in common with other models built by the company. A sample of 27 of the company's planes was taken, and the following model was estimated:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

y = design effort, in millions of worker-hours
x_1 = plane's top speed, in miles per hour
x_2 = plane's weight, in tons
x_3 = percentage of parts in common with other models

The estimated regression coefficients were as follows:

$$b_1 = 0.661 \quad b_2 = 0.065 \quad b_3 = -0.018$$

The total sum of squares and regression sum of squares were found to be as follows:

$$SST = 3.881 \quad \text{and} \quad SSR = 3.549$$

a. Compute and interpret the coefficient of determination.
b. Compute the error sum of squares.
c. Compute the adjusted coefficient of determination.
d. Compute and interpret the coefficient of multiple correlation.

12.20 The following model was fitted to a sample of 30 families in order to explain household milk consumption:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where

y = milk consumption, in quarts per week
x_1 = weekly income, in hundreds of dollars
x_2 = family size

The least squares estimates of the regression parameters were as follows:

$$b_0 = -0.025 \quad b_1 = 0.052 \quad b_2 = 1.14$$

The total sum of squares and regression sum of squares were found to be as follows:

$$SST = 162.1 \quad \text{and} \quad SSR = 88.2$$

a. Compute and interpret the coefficient of determination.
b. Compute the adjusted coefficient of determination.
c. Compute and interpret the coefficient of multiple correlation.

12.21 The following model was fitted to a sample of 25 students using data obtained at the end of their freshman year in college. The aim was to explain students' weight gains:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

y = weight gained, in pounds, during freshman year
x_1 = average number of meals eaten per week
x_2 = average number of hours of exercise per week
x_3 = average number of beers consumed per week

The least squares estimates of the regression parameters were as follows:

$$b_0 = 7.35 \quad b_1 = 0.653 \quad b_2 = -1.345 \quad b_3 = 0.613$$

The regression sum of squares and error sum of squares were found to be as follows:

$$SSR = 79.2 \quad \text{and} \quad SSE = 45.9$$

a. Compute and interpret the coefficient of determination.
b. Compute the adjusted coefficient of determination.
c. Compute and interpret the coefficient of multiple correlation.

12.22 Refer to the savings and loan association data given in Table 12.1.

a. Estimate, by least squares, the regression of profit margin on number of offices.
b. Estimate, by least squares, the regression of net revenues on number of offices.
c. Estimate, by least squares, the regression of profit margin on net revenues.
d. Estimate, by least squares, the regression of number of offices on net revenues.

12.4 Confidence Intervals and Hypothesis Tests for Individual Regression Coefficients

In Section 12.2 we developed and discussed the point estimators for the parameters of the multiple regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

Now, we will develop confidence intervals and tests of hypotheses for the estimated regression coefficients. These confidence intervals and hypothesis tests depend on the variance of the coefficients and the probability distribution of the coefficients. In Section 11.5 we showed that the simple regression coefficient is a linear function of the dependent variable, Y. Multiple regression coefficients, denoted by b_j, are also linear functions of the dependent variable, Y, but the algebra is somewhat more complex and is not presented here. In the previous multiple regression equation, we see that the dependent variable, Y, is a linear function of the X variables plus the random error, ε. For a given set of X terms the function

$$\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki}$$

is actually a constant. We also know from Chapters 4 and 5 that adding a constant to a random variable ε results in the random variable Y having the same probability distribution and variance as the original random variable ε. As a result, the dependent variable, Y, has the same normal distribution and variance as the error term, ε. Then it follows that the regression coefficients, b_j—which are linear functions of Y—also have a normal distribution, and their variance can be derived by using the linear relationship between the regression coefficients and the dependent variable. This computation would follow the same process as used for simple regression in Section 11.5, but the algebra is more complex.

Based on the linear relationship between the coefficients and Y, we know that the coefficient estimates are normally distributed if the model error, ε, is normally distributed. Because of the central limit theorem, we generally find that the coefficient estimates are approximately normally distributed even if ε is not normally distributed. Thus, the hypothesis tests and confidence intervals we develop are not seriously affected by departures from normality in the distribution of the error terms.

We can think of the error term, ε, in the population regression model as including the combined influences on the dependent variable of a multitude of factors not included in the list of independent variables. These factors individually may not have an important influence, but in combination their effect can be important. The fact that the error term is made up of a large number of components whose effects are random provides an intuitive argument for assuming that the coefficient errors are also normally distributed.

As we have seen previously, the coefficient estimators, b_j, are linear functions of Y, and the predicted value of Y is a linear function of the regression coefficient estimators. However, these relationships can sometimes cause interpretation problems. Thus, we will spend time gaining important insights into the variance computations. If we do not understand how the variances are computed, we will not be able to adequately understand hypothesis tests and confidence intervals.

The variance of a coefficient estimate is affected by the sample size, the spread of the X variables, the correlations between the independent variables, and the model error term. Thus, these correlations affect both confidence intervals and tests of hypotheses. Previously, we saw how the correlations between the independent variables influence the coefficient estimators. These correlations between independent variables also increase the variance of the coefficient estimators. An important conclusion is that the variance of the coefficient estimators, in addition to the coefficient estimators, is conditional on the entire set of independent variables in the regression model.

The previous discussion under three-dimensional graphing emphasized the complex effects of several variables on the coefficient variance. As the relationships between independent variables become stronger, estimates of coefficients become more unstable—that is, they have higher variance. The following discussion provides a more formal discussion of these complexities. To obtain good coefficient estimates—those that are low in variance—you should seek a wide range for the independent variables, choose independent variables that are not strongly related to each other, and find a model that is close to all data points. The reality of applied statistical work in business and economics is that we often must use data that are less than ideal, such as the data for the savings and loan example. But by knowing the effects discussed here, we can make good judgments about the applicability of our models.

To gain some understanding of the effect of independent variable correlations, we consider the variance estimators from the estimated multiple regression model with two predictor variables:

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$$

The coefficient variance estimators are

$$s_{b_1}^2 = \frac{s_e^2}{(n-1)s_{x_1}^2(1-r_{x_1x_2}^2)} \qquad (12.16)$$

$$s_{b_2}^2 = \frac{s_e^2}{(n-1)s_{x_2}^2(1-r_{x_1x_2}^2)} \qquad (12.17)$$

and the square roots of these variance estimators, s_{b_1} and s_{b_2}, are called the *coefficient standard errors.*

The variance of the coefficient estimators increases directly with the distance of the points from the line, measured by s_e^2, the estimated error variance. In addition, a wider spread of the independent variable values—measured by $s_{x_1}^2$ or by $s_{x_2}^2$—decreases the coefficient variance. Recall that these results also apply for simple regression coefficient estimators. We also see that the variance of the coefficient estimators increases with increases in the correlation between the independent variables in the model. As the correlation increases between two independent variables, it becomes more difficult to separate the effect of the individual variables for predicting the dependent variables. As the number of independent variables in a model increases, the influences on the coefficient variance continue to be important, but the algebraic structure becomes very complex and is not presented here. The correlation effect leads to the result that coefficient variance estimators are conditional on the other independent variables in the model. Recall that the actual coefficient estimators are also conditional on the other independent variables in the model, again because of the effect of correlations between the independent variables.

The basis for inference about population regression coefficients is summarized next. We are typically more interested in the regression coefficients β_j than in the constant or intercept β_0. Thus, we concentrate on the former, noting that inference about the latter proceeds along similar lines.

Basis for Inference about the Population Regression Parameters

Let the population regression model be as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

Let $b_0, b_1, \ldots, b_K$ be the least squares estimates of the population parameters and $s_{b_0}, s_{b_1}, \ldots, s_{b_K}$ be the estimated standard deviations of the least squares estimators. Then, if the standard regression assumptions hold and if the error terms, ε_i, are normally distributed,

$$t_{b_j} = \frac{b_j - \beta_j}{s_{b_j}} \quad (j = 1, 2, \ldots, K) \qquad (12.18)$$

is distributed as a Student's t distribution with $(n - K - 1)$ degrees of freedom.

Confidence Intervals

Confidence intervals for the β_j can be derived by using Equation 12.19.

Confidence Intervals for Regression Coefficients

If the population regression errors, ε_i, are normally distributed and the standard regression assumptions hold, the $100(1-\alpha)\%$ two-sided **confidence intervals for the regression coefficients**, β_j, are given by

$$b_j - t_{n-K-1,\alpha/2}s_{b_j} < \beta_j < b_j + t_{n-K-1,\alpha/2}s_{b_j} \qquad (12.19)$$

where $t_{n-K-1, \alpha/2}$ is the number for which

$$P(t_{n-K-1} > t_{n-K-1, \alpha/2}) = \frac{\alpha}{2}$$

and the random variable t_{n-K-1} follows a Student's t distribution with $(n - K - 1)$ degrees of freedom.

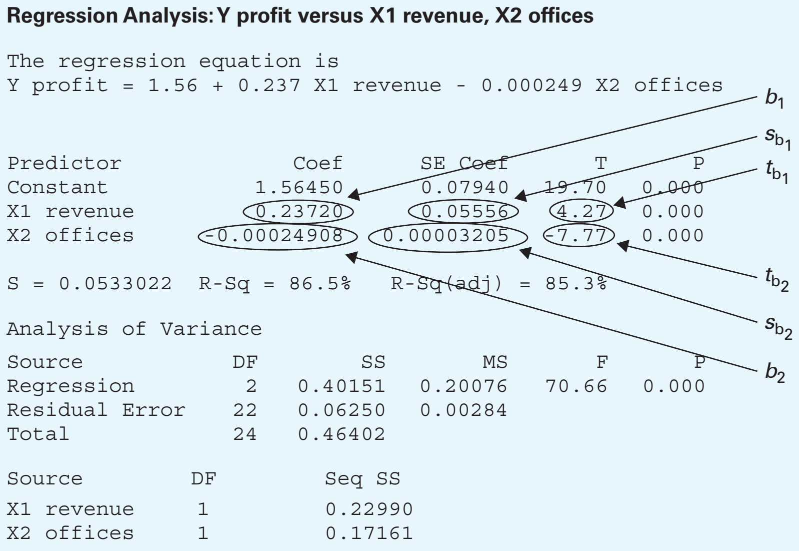
Example 12.4 Developing the Savings and Loan Model (Confidence Interval Estimation)

We have been asked to determine confidence intervals for the coefficients of the savings and loan regression model developed in Example 12.3.

Solution The Minitab regression output for the savings and loan regression model is shown in Figure 12.8. The coefficient estimators and their standard deviations for the revenue, b_1, and number of offices, b_2, predictor variables are computed as follows:

$$b_1 = 0.2372, \quad s_{b_1} = 0.0556; \quad b_2 = -0.000249 \quad \text{and} \quad s_{b_2} = 0.00003205$$

Figure 12.8 Savings and Loan Regression: Minitab Output



Regression Analysis: Y profit versus X1 revenue, X2 offices

```
The regression equation is
Y profit = 1.56 + 0.237 X1 revenue - 0.000249 X2 offices


Predictor              Coef      SE Coef      T      P
Constant            1.56450      0.07940  19.70  0.000
X1 revenue          0.23720      0.05556   4.27  0.000
X2 offices      -0.00024908   0.00003205  -7.77  0.000

S = 0.0533022  R-Sq = 86.5%   R-Sq(adj) = 85.3%

Analysis of Variance

Source            DF       SS       MS      F      P
Regression         2  0.40151  0.20076  70.66  0.000
Residual Error    22  0.06250  0.00284
Total             24  0.46402

Source         DF         Seq SS
X1 revenue      1         0.22990
X2 offices      1         0.17161
```

Thus, we see that the standard deviation of the sampling distribution of the least squares estimator for β_1 is estimated as 0.05556 and for β_2 is estimated as 0.00003205.

To obtain the 99% confidence intervals for β_1 and β_2, we use the Student's t value from Appendix Table 8.

$$t_{n-K-1, \alpha/2} = t_{22, 0.005} = 2.819$$

Using these results, we find that the 99% coefficient confidence interval for β_1 is

$$0.237 - (2.819)(0.05556) < \beta_1 < 0.237 + (2.819)(0.05556)$$

or

$$0.080 < \beta_1 < 0.394$$

Thus, the 99% confidence interval for the expected increase in the savings and loan profit margin resulting from a one-unit increase in net revenue per dollar, given a fixed number of offices, runs from 0.080 to 0.394. The 99% coefficient confidence interval for β_2 is

$$-0.000249 - (2.819)(0.0000320) < \beta_2 < -0.000249 + (2.819)(0.0000320)$$

or

$$-0.000339 < \beta_2 < -0.000159$$

Therefore, we see that the 99% confidence interval for the expected decrease in the profit margin resulting from an increase of 1,000 offices, for a fixed level of net revenue per dollar, runs from 0.159 to 0.339.

Tests of Hypotheses

Tests of hypotheses for regression coefficients can be developed using the coefficient variance estimates. Of particular interest is the hypothesis test

$$H_0 : \beta_j = 0$$

which is frequently used to determine if a specific independent variable is conditionally important in a multiple regression model.

Tests of Hypotheses for the Regression Coefficients

If the regression errors, ε_i, are normally distributed and the standard regression assumptions hold, then the following hypothesis tests have significance level α:

1. To test either null hypothesis

   $$H_0 : \beta_j = \beta^* \quad \text{or} \quad H_0 : \beta_j \leq \beta^*$$

   against the alternative

   $$H_1 : \beta_j > \beta^*$$

   the decision rule is as follows:

   $$\text{reject } H_0 \text{ if } \frac{b_j - \beta^*}{s_{b_j}} > t_{n-K-1,\alpha} \tag{12.20}$$

2. To test either null hypothesis

   $$H_0 : \beta_j = \beta^* \quad \text{or} \quad H_0 : \beta_j \geq \beta^*$$

   against the alternative

   $$H_1 : \beta_j < \beta^*$$

   the decision rule is as follows:

   $$\text{reject } H_0 \text{ if } \frac{b_j - \beta^*}{s_{b_j}} < -t_{n-K-1,\alpha} \tag{12.21}$$

3. To test the null hypothesis

   $$H_0 : \beta_j = \beta^*$$

   against the two-sided alternative

   $$H_1 : \beta_j \neq \beta^*$$

   the decision rule is as follows:

   $$\text{reject } H_0 \text{ if } \frac{b_j - \beta^*}{s_{b_j}} > t_{n-K-1,\alpha/2} \quad \text{or} \quad \frac{b_j - \beta^*}{s_{b_j}} < -t_{n-K-1,\alpha/2} \tag{12.22}$$

Many analysts argue that if we cannot reject the conditional hypothesis that the coefficient is 0, then we must conclude that the variable should not be included in the regression model. The Student's t statistic for this two-tailed test is typically computed in most regression programs and is printed next to the coefficient variance estimate; in addition, the p-value for the hypothesis test is typically included. These are shown in the Minitab output in Figure 12.8. Using the printed Student's t statistic or the p-value, we can immediately conclude whether or not a particular predictor variable is conditionally significant, given the other variables in the regression model.

There are clearly other procedures for deciding if an independent variable should be included in a regression model. We see that the preceding selection procedure ignores Type II error—the population coefficient is not equal to 0, but we fail to reject the null hypothesis that it is equal to 0. This is a particular problem when a model based on economic or another theory that is carefully specified to include certain independent variables. Then, because of a large error, ε, or correlations between independent variables, or both, we cannot reject the hypothesis that the coefficient is 0. In this case many analysts will include the independent variable in the model because the original model specification based on economic theory or experience is believed to dominate. This is a difficult issue and requires good judgment based on both statistical results and theory concerning the underlying relationship being modeled.

Example 12.5 Developing the Savings and Loan Model (Coefficient Hypothesis Tests)

We have been asked to determine if the coefficients in the savings and loan regression model are conditionally significant predictors of profit margin.

Solution The hypothesis test for this question will use the Minitab regression results shown in Figure 12.8. First, we wish to determine if the variable net revenue per dollar has a significant effect on increasing profit margin, conditional on or controlling for the effect of the variable number of offices. The null hypothesis is

$$H_0 : \beta_1 = 0$$

versus the alternative hypothesis

$$H_1 : \beta_1 > 0$$

The test can be performed by computing the Student's t statistic associated with the coefficient, given H_0:

$$t_{b_1} = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{0.237 - 0}{0.05556} = 4.27$$

From the Student's t table, Appendix Table 8, we can determine that the critical value—for $\alpha = 0.005$– for the Student's t statistic is as follows:

$$t_{22,0.005} = 2.819$$

Figure 12.8 also indicates that the p-value for the null hypothesis test

$$H_0 : \beta_1 = 0$$

versus the alternative hypothesis

$$H_1 : \beta_1 \neq 0$$

is less than 0.005. Based on this evidence, we reject H_0 and accept H_1 and conclude that net revenue per dollar is a statistically significant predictor of increased profit margin for savings and loans, given that we have controlled for the effect of the number of offices.

Similarly, we can determine if the total number of offices has a significant effect on reducing profit margins. The null hypothesis is

$$H_0 : \beta_2 = 0$$

versus the alternative hypothesis

$$H_1 : \beta_2 < 0$$

The test can be performed by computing the Student's t statistic associated with the coefficient, given H_0:

$$t_{b_2} = \frac{b_2 - \beta_2}{s_{b_2}} = \frac{-0.000249 - 0}{0.0000320} = -7.77$$

From Appendix Table 8 we find that the critical value for the Student's t statistic is as follows:

$$t_{22,\,0.005} = -2.819$$

Figure 12.8 also indicates that the p-value for the null hypothesis test

$$H_0 : \beta_2 = 0$$

versus the alternative hypothesis

$$H_1 : \beta_2 \neq 0$$

is less than 0.005. Based on this evidence, we reject H_0 and accept H_1 and conclude that number of offices is a statistically significant predictor of lower profit margin for savings and loans, given that we have controlled for the effect of net revenue per dollar.

It is important to emphasize that both of the hypothesis tests are based on the particular set of variables included in the regression model. If, for example, additional predictor variables were included, then these tests would no longer be valid. With additional variables in the model the coefficient estimates and their estimated standard deviations would be different, and, thus, the Student's t statistics would also be different.

Note that in the Minitab regression output for this problem, shown in Figure 12.8, the Student's t statistic for the null hypothesis—$H_0 : \beta_j = 0$—is computed as the ratio of the estimated coefficient divided by the estimated coefficient standard error—contained in the two columns to the left of the Student's t. The probability, or p-value, for the two-tailed hypothesis test—$H_j : \beta_j \neq 0$—is also displayed. Thus, an analyst can perform these hypothesis tests directly by examining the multiple regression output. The Student's t and the p-value are computed in every modern statistical package. Most analysts routinely look for these test results as they examine regression output from a computer statistical package.

Example 12.6 Factors Affecting Property Tax Rate (Analysis of Regression Coefficients)

A group of city managers commissioned a study to determine the factors that influence urban property-tax rates for cities with populations between 100,000 and 200,000.

Solution Using a sample of 20 U.S. cities, the following regression model was estimated:

$$\hat{y} = 1.79 + \underset{(0.000139)}{0.000567x_1} + \underset{(0.0082)}{0.0183x_2} - \underset{(0.000446)}{0.000191x_3}$$

$$R^2 = 0.71 \qquad n = 20$$

where

y = effective property tax rate (actual levies divided by market value of the tax base)
x_1 = number of housing units per square mile
x_2 = percentage of total city revenue represented by grants from state and federal governments
x_3 = median per capita personal income, in dollars

The numbers in parentheses under the coefficients are the estimated coefficient standard errors.

The preceding presentation of the regression equation and variable definition provides a good format for displaying the results of a regression analysis model. The results indicate that the conditional estimates of the effects of the three predictor variables are as follows:

1. An increase of one housing unit per square mile increases the effective property tax rate by 0.000567. Note that property tax rates are typically expressed in terms of dollars per $1,000 of assessed property value. Thus, an increase of 0.000567 indicates that property tax rates are higher by $0.567 per $1,000 of assessed property value.
2. An increase of 1% of the total city revenue from state and federal grants increases the effective tax rate by 0.0183.
3. An increase of $1 in median per capita personal income leads to an expected decrease in the effective tax rate by 0.000191. Note that the ratio of 0.000191 divided by 0.000446 gives a t value less than 2.

We emphasize again that these coefficient estimates are valid only for a model with all three predictor variables included.

To better understand the accuracy of these effects, we construct conditional 95% confidence intervals. For the estimated regression model there are $(20 - 3 - 1) = 16$ degrees of freedom for error. Thus, the Student's t statistic for computing confidence intervals is, from the Appendix, $t_{16,0.025} = 2.12$. The format for confidence intervals is as follows:

$$b_j - t_{n-K-1,\,\alpha/2}s_{bj} < \beta_j < b_j + t_{n-K-1,\,\alpha/2}s_{bj}$$

Thus, the coefficient for the number of housing units per square mile has a 95% confidence interval of

$$0.000567 - (2.12)(0.000139) < \beta_1 < 0.000567 + (2.12)(0.000139)$$
$$0.000272 < \beta_1 < 0.000862$$

The coefficient for the percentage of revenue represented by grants has a 95% confidence interval of

$$0.0183 - (2.12)(0.0082) < \beta_2 < 0.0183 + (2.12)(0.0082)$$
$$0.0009 < \beta_2 < 0.0357$$

Finally, the coefficient for median per capita personal income has a 95% confidence interval of

$$-0.000191 - (2.12)(0.000446) < \beta_3 < -0.000191 + (2.12)(0.000446)$$
$$-0.001137 < \beta_3 < 0.000755$$

Again, we emphasize that these intervals are conditional on all three predictor variables being included in the model.

We see that the 95% confidence interval for β_3 includes 0, and, thus, we could not reject the two-tailed hypothesis that this coefficient is 0. Based on this confidence interval, we conclude that X_3 is not a statistically significant predictor variable in the multiple regression model. However, the confidence intervals for the other two variables do not include 0, and, thus, we conclude that they are statistically significant.

Example 12.7 Effects of Fiscal Factors on Housing Prices (Regression Model Coefficient Estimation)

Northern City, Minnesota, was interested in the effect of local property development on the market price of houses in the city. Northern City is one of many small, nonmetropolitan, midwestern cities with populations in the range from 6,000 to 40,000. One of the objectives was to determine how increased commercial property development would influence the value of local housing. Data are stored in the data file **Citydatr**.

Solution To answer this question, data were collected from a number of cities and used to construct a regression model that estimates the effect of key variables on housing price. For this study the following variables were obtained for each city:

$Y\,(\text{hseval}) =$ mean market price for houses in the city
$X_1\,(\text{sizehse}) =$ mean number of rooms in houses
$X_2\,(\text{incom72}) =$ mean household income
$X_3\,(\text{taxrate}) =$ tax rate per thousand dollars of assessed value for houses
$X_4\,(\text{Comper}) =$ percentage of taxable property that is commercial property

The multiple regression output, prepared using Minitab, is shown in Figure 12.9. The coefficient for the mean number of rooms in city houses is 7.878, with a coefficient standard deviation of 1.809. In this study housing values are in units of $1,000, with a mean of $21,000 over all cities. Thus, if the mean number of rooms in a city's houses was larger by 1.0, then the mean price would be larger by $7,878. The resulting Student's t statistic is 4.35 and the p-value is 0.000. Thus, the conditional hypothesis that this coefficient is equal to 0 is rejected. The same result occurs for the income and tax rate variables. The incom72 variable is in units of dollars, and, thus, if a city's mean income is higher by $1,000, the coefficient of 0.003666 indicates that mean housing price will be $3,666 higher. If the tax rate increases by 1%, mean housing price is reduced by $1,718. We see that the regression analysis leads to the conclusion that each of these three variables is a significant predictor of the mean house price in the cities included in this study. However, we see that the coefficient for the percent of commercial property, Comper, is -10.614, with a coefficient standard deviation of 6.491, resulting in a Student's t statistic equal to -1.64. Note that here is an important area for judgment. The coefficient would have a single-tail p-value of 0.053 or a two-tailed p-value of 0.106. Thus, it appears to have some effect in reducing the mean price of houses. Given that the effects of house size, income, and tax rate on the market price for houses have been included, we see that the percent of commercial property does not increase housing prices. Thus, the argument that the market value of houses will increase if more commercial property is developed is not supported by this analysis. That conclusion is true only for a model that includes these four predictor variables. Note also that the values of $R^2 = 47.4\%$ and s_e (standard error of the regression) $= 3.677$ are included in the regression output.

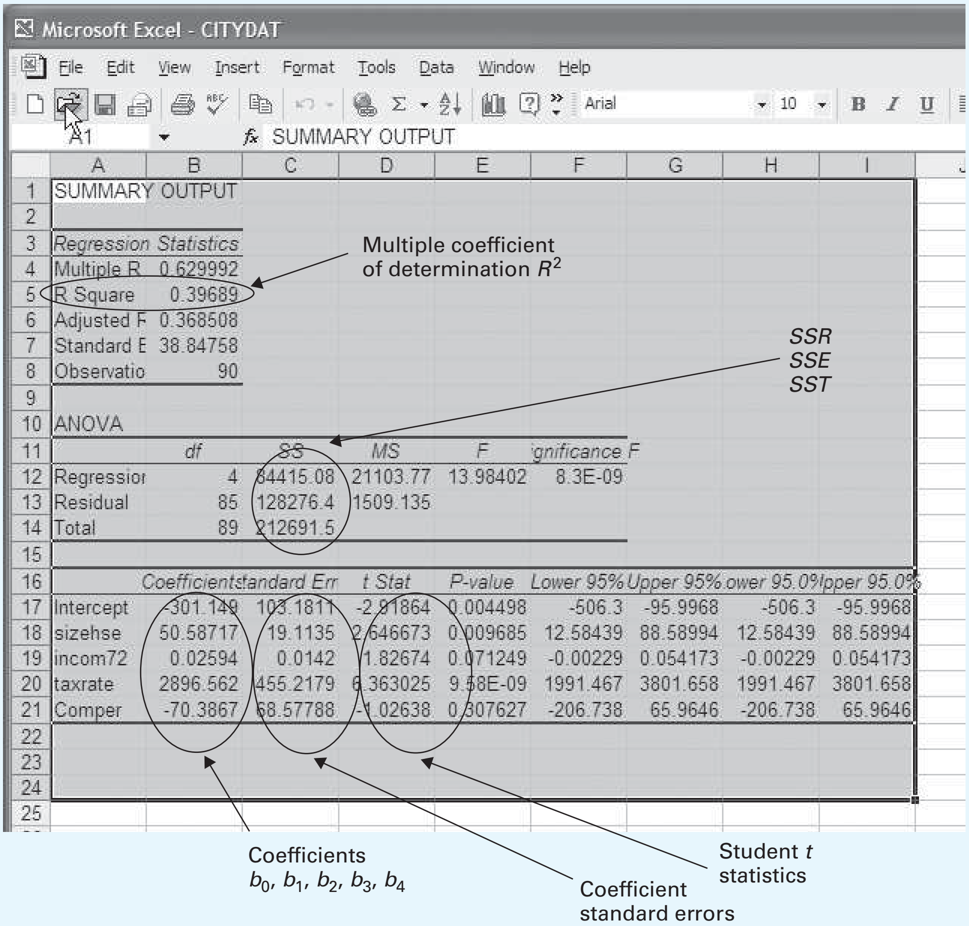
Figure 12.9 Housing Price Regression Model (Minitab Output)

Regression Analysis: hseval versus sizehse, income72, taxrate, Comper

```
The regression equation is
hseval = -28.1 + 7.88 sizehse + 0.00367 incom72 - 172 taxrate -10.6 Comper

Predictor         Coef     SE Coef        T       P
Constant       -28.075       9.766    -2.87   0.005
sizehse          7.878       1.809     4.35   0.000
incom72       0.003666    0.001344     2.73   0.008
taxrate        -171.80       43.09    -3.99   0.000
Comper         -10.614       6.491    -1.64   0.106


S = 3.67686    R-Sq = 47.4%   R-Sq(adj) = 45.0%

Analysis of Variance

Source            DF        SS        MS       F       P
Regression         4   1037.49    259.37   19.19   0.000
Residual Error    85   1149.14     13.52
Total             89   2186.63
```

The advocates of increased commercial development also claimed that increasing the amount of commercial property would decrease the taxes paid on owner-occupied houses. This claim was tested using the regression output in Figure 12.10, prepared using Excel. The coefficient estimators and their standard errors are indicated. The Student's *t* statistics for the size of house and the tax-rate coefficients are 2.65 and 6.36, indicating that these variables are important predictors. The Student's *t* statistic for income is 1.83, with a *p*-value of 0.07 for a two-tailed test. Thus, income has some influence as a predictor, but its effect is not as strong as the previous two variables. Again, we see a place for good judgment that considers the problem context. The conditional hypothesis that increased commercial property decreases taxes on owner-occupied houses can be tested using the conditional Student's *t* statistic for the variable "Comper" in the regression output. The conditional Student's *t* statistic is −1.03, with a *p*-value of 0.308. Thus, the hypothesis that increased commercial property does not decrease house taxes cannot be rejected. There is no evidence from this analysis that house taxes would be lowered if there was additional commercial development.

Figure 12.10 House-Tax Regression Model (Excel Output)


Microsoft Excel - CITYDAT
File
Edit
View
Insert
Format
Tools
Data
Window
Help
Arial
10
B
I
U
A1
fx
SUMMARY OUTPUT
Multiple coefficient of determination R^2
SSR
SSE
SST
Coefficients b_0, b_1, b_2, b_3, b_4
Coefficient standard errors
Student *t* statistics

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.629992							
5	R Square	0.39689							
6	Adjusted F	0.368508							
7	Standard E	38.84758							
8	Observatio	90							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regressior	4	84415.08	21103.77	13.98402	8.3E-09			
13	Residual	85	128276.4	1509.135					
14	Total	89	212691.5						
15									
16		*Coefficients*	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*lpper 95.0%*
17	Intercept	-301.149	103.1811	-2.91864	0.004498	-506.3	-95.9968	-506.3	-95.9968
18	sizehse	50.58717	19.1135	2.646673	0.009685	12.58439	88.58994	12.58439	88.58994
19	incom72	0.02594	0.0142	1.82674	0.071249	-0.00229	0.054173	-0.00229	0.054173
20	taxrate	2896.562	455.2179	6.363025	9.58E-09	1991.467	3801.658	1991.467	3801.658
21	Comper	-70.3867	68.57788	-1.02638	0.307627	-206.738	65.9646	-206.738	65.9646

Based on the regression analyses performed in this study, the consultants concluded that there was no evidence that increased commercial property would either increase the market value of houses or lower the property taxes for a house.

Exercises

Visit **www.MyStatLab.com** or **www.pearsonhighered.com/newbold** to access the data files.

Basic Exercises

12.23 The following are results from a regression model analysis:

$$\hat{y} = 1.50 + \underset{(2.1)}{4.8x_1} + \underset{(3.7)}{6.9x_2} - \underset{(2.8)}{7.2x_3}$$

$$R^2 = 0.71 \qquad n = 24$$

The numbers below the coefficient estimates are the sample standard errors of the coefficient estimates.

a. Compute two-sided 95% confidence intervals for the three regression slope coefficients.

b. For each of the slope coefficients, test the hypothesis

$$H_0 : \beta_j = 0$$

12.24 The following are results from a regression model analysis:

$$\hat{y} = 2.50 + \underset{(3.1)}{6.8x_1} + \underset{(3.7)}{6.9x_2} - \underset{(3.2)}{7.2x_3}$$

$$R^2 = 0.85 \qquad n = 34$$

The numbers below the coefficient estimates are the estimated coefficient standard errors.

a. Compute two-sided 95% confidence intervals for the three regression slope coefficients.

b. For each of the slope coefficients test the hypothesis

$$H_0 : \beta_j = 0$$

12.25 The following are results from a regression model analysis:

$$\hat{y} = -101.50 + \underset{(12.1)}{34.8x_1} + \underset{(23.7)}{56.9x_2} - \underset{(32.8)}{57.2x_3}$$

$$R^2 = 0.71 \qquad n = 65$$

The numbers in parentheses under the coefficients are the estimated coefficient standard errors.

a. Compute two-sided 95% confidence intervals for the three regression slope coefficients.

b. For each of the slope coefficients test the hypothesis

$$H_0 : \beta_j = 0$$

12.26 The following are results from a regression model analysis:

$$\hat{y} = -9.50 + \underset{(7.1)}{17.8x_1} + \underset{(13.7)}{26.9x_2} - \underset{(3.8)}{9.2x_3}$$

$$R^2 = 0.71 \qquad n = 39$$

The numbers in parentheses under the coefficients are the estimated coefficient standard errors.

a. Compute two-sided 95% confidence intervals for the three regression slope coefficients.

b. For each of the slope coefficients test the hypothesis

$$H_0 : \beta_j = 0$$

Application Exercises

12.27 An aircraft company wanted to predict the number of worker-hours necessary to finish the design of a new plane. Relevant explanatory variables were thought to be the plane's top speed, its weight, and the number of parts it had in common with other models built by the company. A sample of 27 of the company's planes was taken, and the following model was estimated:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

y = design effort, in millions of worker-hours
x_1 = plane's top speed, in miles per hour
x_2 = plane's weight, in tons
x_3 = percentage of parts in common with other models

The estimated regression coefficients were as follows:

$$b_1 = 0.661 \quad b_2 = 0.065 \quad b_3 = -0.018$$

The estimated standard errors were as follows:

$$s_{b_1} = 0.099 \quad s_{b_2} = 0.032 \quad s_{b_3} = 0.0023$$

a. Find 90% and 95% confidence intervals for β_1.

b. Find 95% and 99% confidence intervals for β_2.

c. Test against a two-sided alternative the null hypothesis that, all else being equal, the plane's weight has no linear influence on its design effort.

d. The error sum of squares for this regression was 0.332. Using the same data, a simple linear regression of design effort on the percentage of common parts was fitted, yielding an error sum of squares of 3.311. Test, at the 1% level, the null hypothesis that, taken together, the variable's top speed and weight contribute nothing in a linear sense to explaining the changes in the variable, design effort, given that the variable percentage of common parts is also used as an explanatory variable.

12.28 The following model was fitted to a sample of 30 families in order to explain household milk consumption:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where

y = milk consumption, in quarts per week
x_1 = weekly income, in hundreds of dollars
x_2 = family size

The least squares estimates of the regression parameters were as follows:

$$b_0 = -0.025 \quad b_1 = 0.052 \quad b_2 = 1.14$$

The estimated standard errors were as follows:

$$s_{b_1} = 0.023 \quad s_{b_2} = 0.35$$

a. Test, against the appropriate one-sided alternative, the null hypothesis that, for fixed family size, milk consumption does not depend linearly on income.

b. Find 90%, 95%, and 99% confidence intervals for β_2.

12.29 The following model was fitted to a sample of 25 students using data obtained at the end of their freshman year in college. The aim was to explain students' weight gains:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

y = weight gained, in pounds, during freshman year
x_1 = average number of meals eaten per week
x_2 = average number of hours of exercise per week
x_3 = average number of beers consumed per week

The least squares estimates of the regression parameters were as follows:

$$b_0 = 7.35 \quad b_1 = 0.653 \quad b_2 = -1.345 \quad b_3 = 0.613$$

The estimated standard errors were as follows:

$$s_{b_1} = 0.189 \quad s_{b_2} = 0.565 \quad s_{b_3} = 0.243$$

a. Test, against the appropriate one-sided alternative, the null hypothesis that, all else being equal, hours of exercise do not linearly influence weight gain.
b. Test, against the appropriate one-sided alternative, the null hypothesis that, all else being equal, beer consumption does not linearly influence weight gain.
c. Find 90%, 95%, and 99% confidence intervals for β_1.

12.30 Refer to the data of Example 12.6.

a. Test, against a two-sided alternative, the null hypothesis that, all else being equal, median per capita personal income has no influence on the effective property tax rate.
b. Test the null hypothesis that, taken together, the three independent variables do not linearly influence the effective property tax rate.

12.31 Refer to the data of Example 12.7 with the data file **Citydatr**.

a. Find 95% and 99% confidence intervals for the expected change in the market price for houses resulting from a one-unit increase in the mean number of rooms when the values of all other independent variables remain unchanged.
b. Test the null hypothesis that, all else being equal, mean household income does not influence the market price against the alternative that the higher the mean household income, the higher the market price.

12.32 In a study of revenue generated by national lotteries, the following regression equation was fitted to data from 29 countries with lotteries:

$$y = -31.323 + \underset{(0.00755)}{0.4045x_1} + \underset{(0.3107)}{0.8772x_2} - \underset{(263.88)}{365.01x_3} - \underset{(3.4520)}{9.9298x_4}$$

$R^2 = .51$

where

y = dollars of net revenue per capita per year generated by the lottery
x_1 = mean per capita personal income of the country
x_2 = number of hotel, motel, inn, and resort rooms per thousand persons in the country
x_3 = spendable revenue per capita per year generated by pari-mutuel betting, racing, and other legalized gambling
x_4 = percentage of the nation's border contiguous with a state or states with a lottery

The numbers in parentheses under the coefficients are the estimated coefficient standard errors.

a. Interpret the estimated coefficient on x_1.
b. Find and interpret a 95% confidence interval for the coefficient on x_2 in the population regression.
c. Test the null hypothesis that the coefficient on x_3 in the population regression is 0 against the alternative that this coefficient is negative. Interpret your findings.

12.33 A study was conducted to determine whether certain features could be used to explain variability in the prices of furnaces. For a sample of 19 furnaces, the following regression was estimated:

$$y = -68.236 + \underset{(0.005)}{0.0023x_1} + \underset{(8.992)}{19.729x_2} + \underset{(3.082)}{7.653x_3} \quad R^2 = 0.84$$

where

y = price, in dollars
x_1 = rating of furnace, in BTU per hour
x_2 = energy efficiency ratio
x_3 = number of settings

The numbers in parentheses under the coefficients are the estimated coefficient standard errors.

a. Find a 95% confidence interval for the expected increase in price resulting from an additional setting when the values of the rating and the energy efficiency ratio remain fixed.
b. Test the null hypothesis that, all else being equal, the energy efficiency ratio of furnaces does not affect their price against the alternative that the higher the energy efficiency ratio, the higher the price.

12.34 In a study of differences in levels of community demand for firefighters, the following sample regression was obtained, based on data from 39 towns in Maryland:

$$y = -0.00232 - \underset{(0.00010)}{0.00024x_1} - \underset{(0.000018)}{0.00002x_2} + \underset{(0.00012)}{0.00034x_3}$$
$$+ \underset{(0.77954)}{0.48122x_4} + \underset{(0.01172)}{0.04950x_5} - \underset{(0.00005)}{0.00010x_6} + \underset{(0.00306)}{0.00645x_7}$$

$\bar{R}^2 = 0.3572$

where

y = number of full-time firefighters per capita
x_1 = maximum base salary of firefighters, in thousands of dollars
x_2 = percentage of population
x_3 = estimated per capita income, in thousands of dollars
x_4 = population density
x_5 = amount of intergovernmental grants per capita, in thousands of dollars
x_6 = number of miles from the regional city
x_7 = percentage of the population that is male and between 12 and 21 years of age

The numbers in parentheses under the coefficients are the estimated coefficient standard errors.

a. Find and interpret a 99% confidence interval for β_5.
b. Test, against a two-sided alternative, the null hypothesis that β_4 is 0, and interpret your result.
c. Test, against a two-sided alternative, the null hypothesis that β_7 is 0, and interpret your result.

12.5 Tests on Regression Coefficients

In the previous section we showed how a conditional hypothesis test can be conducted to determine if a specific variable coefficient is conditionally significant in a regression model. There are, however, situations where we are interested in the effect of the combination of several variables. For example, in a model that predicts quantity sold, we might be interested in the combined effect of both the seller's price and the competitor's price. In other cases we might be interested in knowing if the combination of all variables is a useful predictor of the dependent variable.

Tests on All Coefficients

First, we present hypothesis tests to determine if sets of several coefficients are all simultaneously equal to 0. Consider again the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1 + \cdots + \beta_K x_K + \varepsilon$$

We begin by considering the null hypothesis that all the coefficients are simultaneously equal to zero:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_K = 0$$

Accepting this hypothesis would lead us to conclude that none of the predictor variables in the regression model is statistically significant and, thus, that they provide no useful information. If this were to occur, then we would need to go back to the model-specification process and develop a new set of predictor variables. Fortunately, in most applied regression situations this hypothesis is rejected because the specification process usually leads to identification of at least one significant predictor variable.

To test this hypothesis, we can use the partitioning of variability developed in Section 12.3:

$$SST = SSR + SSE$$

Recall that *SSR* is the amount of variability explained by the regression and that *SSE* is the amount of unexplained variability. Also recall that the variance of the regression model can be estimated by using the following:

$$s_e^2 = \frac{SSE}{(n - K - 1)}$$

If the null hypothesis that all coefficients are equal to 0 is true, then *the mean square regression,*

$$MSR = \frac{SSR}{K}$$

is also a measure of error with K degrees of freedom. As a result, the ratio

$$\begin{aligned} F &= \frac{SSR/K}{SSE/(n - K - 1)} \\ &= \frac{MSR}{s_e^2} \end{aligned}$$

has an F distribution with K degrees of freedom for the numerator and $(n - K - 1)$ degrees of freedom for the denominator. If the null hypothesis is true, then both the numerator and the denominator provide estimates of the population variance. As noted in Section 11.5, the ratio of independent sample variances from populations with equal population variances follows an F distribution if the populations are normally distributed. The computed value of F is compared with the critical value of F from Appendix Table 9 at a significance level α. If the computed value exceeds the critical value from the table, we reject the null hypothesis and conclude that at least one coefficient is not equal to 0. This test procedure is summarized in Equation 12.23.

Test on All the Coefficients of a Regression Model

Consider the multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \varepsilon$$

To test the null hypothesis

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_K = 0$$

against the alternative hypothesis

$$H_1: \text{at least one } \beta_j \neq 0$$

at a significance level α, we use the decision rule

$$\text{reject } H_0: \text{if} \qquad F_{K,n-K-1} = \frac{MSR}{s_e^2} > F_{K,n-K-1,\alpha} \tag{12.23}$$

where $F_{K,n-K-1,\alpha}$ is the critical value of F from Appendix Table 9 for which

$$P\left(F_{K,n-K-1} > F_{K,n-K-1,\alpha}\right) = \alpha$$

The computed random variable $F_{K,n-K-1}$ follows an F distribution with numerator degrees of freedom K and denominator degrees of freedom $(n - K - 1)$.

Example 12.8 Housing Price Prediction Model (Simultaneous Coefficient Testing)

During the development of the housing price prediction model for Northern City, the analysts wanted to know if there was evidence that the combination of four predictor variables was not a significant predictor of housing price. That is, they wanted to test, at a 99% confidence level, the hypothesis

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

Solution This testing procedure can be illustrated by the housing price regression in Figure 12.9 prepared using the **Citydatr** data file. In the analysis of variance table, the computed F statistic is 19.19, with 4 degrees of freedom for the numerator and 85 degrees of freedom for the denominator. The computation of F is as follows:

$$F = \frac{259.37}{13.52} = 19.184$$

This exceeds the critical value of $F = 3.548$ for $\alpha = 0.01$ from Appendix Table 9. In addition, note that Minitab—and most statistics packages—compute the p-value, which in this example is equal to 0.000. Thus, we would reject the hypothesis that all coefficients are equal to zero.

Test on a Subset of Regression Coefficients

In the previous sections we developed hypothesis tests for individual regression parameters and for all regression parameters taken together. Next, we develop a hypothesis test for a subset of regression parameters, such as the combined price example previously discussed. We use this test to determine if the combined effect of several independent variables is significant in a regression model.

Consider a regression model that contains independent variables designated as X_j and Z_j terms:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \alpha_1 z_1 + \cdots + \alpha_R z_R + \varepsilon$$

and the null hypothesis to be tested is as follows:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_R = 0 \text{ given } \beta_j \neq 0, j = 1, \ldots, K$$

If H_0 is true, then the Z_j variables should not be included in the regression model because they provide nothing further to explain the behavior of the dependent variable beyond what the X_j variables provided. The procedure for performing this test is summarized in Equation 12.24, following a detailed discussion of the testing procedure.

The test is conducted by comparing the error sum of squares, SSE, from the complete regression model, which includes both the X and the Z variables, with the $SSE(R)$ from a restricted model that includes only the X variables. First, we run a regression on the complete regression model and obtain the error sum of squares, designated as SSE. Next, we run the restricted regression, which excludes the Z variables (note that the coefficients α_j are all restricted to values of 0 in this regression):

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \varepsilon^*$$

From this regression we obtain the restricted error sum of squares, designated as $SSE(R)$. Then we compute the F statistic with r degrees of freedom for the numerator, where r is the number of variables removed simultaneously from the restricted model and there are $(n - K - R - 1)$ degrees of freedom for the denominator, the degrees of freedom for error in the model that includes both the X and the Z independent variables. The F statistic is

$$F = \frac{(SSE(R) - SSE)/R}{s_e^2}$$

where s_e^2 is the estimated variance of the error for the complete model. This statistic follows an F distribution with R degrees of freedom in the numerator and $(n - K - R - 1)$ degrees of freedom in the denominator. If the computed F is greater than the critical value of F, then the null hypothesis is rejected, and we conclude that the Z variables as a set should be included in the model. Note that this test does not imply that individual Z variables should not be excluded by, for example, using the Student's t test discussed previously. In addition, the test for all Z's does not imply that a subset of the Z variables cannot be excluded by using this test procedure with a different subset of Z variables.

Test on a Subset of the Regression Parameters

Given a regression model with the independent variables partitioned into X and Z subsets,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \alpha_1 z_1 + \cdots + \alpha_R z_R + \varepsilon$$

To test the null hypothesis

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \cdots = \alpha_R = 0$$

which states that the regression parameters in a particular subset are simultaneously equal to 0, against the alternative hypothesis

$$H_1 : \text{At least one } \alpha_j \neq 0 \; (j = 1, \ldots, R)$$

We compare the error sum of squares for the complete model with the error sum of squares for the restricted model. First, run a regression for the complete model, which includes all independent variables, and obtain the error sum of squares, SSE. Next, run a restricted regression, which excludes the Z variables whose coefficients are the α_i's—the number of variables excluded is R. From this regression obtain the restricted error sum of squares, $SSE(R)$. Then compute the F statistic and apply the decision rule for significance level α:

$$\text{reject } H_0 \text{ if } F = \frac{(SSE(R) - SSE)/R}{s_e^2} > F_{R, n-K-R-1, \alpha} \qquad (12.24)$$

Comparison of F and t Tests

If we used Equation 12.24 with $R = 1$, we could test the hypothesis that a single variable, X_j, does not improve the prediction of the dependent variable, given the other independent variables in the model. Thus, we have the following hypothesis test:

$$H_0 : \beta_j = 0 | \beta_l \neq 0, \quad j \neq l \quad l = 1, \ldots, K$$
$$H_1 : \beta_j \neq 0 | \beta_l \neq 0, \quad j \neq l \quad l = 1, \ldots, K$$

Previously, we saw that this test could also be performed using a Student's t test. Using methods beyond this book, we can show that the corresponding F and t tests provide exactly the same conclusions regarding the hypothesis test for a single variable. In addition, the computed t statistic for the coefficient b_j is equal to the square root of the corresponding computed F statistic. That is,

$$t_{b_j}^2 = F_{x_j}$$

where F_{x_j} is the F statistic computed using Equation 12.24 when variable x_j is excluded from the model and, thus, $R = 1$. We show this numerical result in Example 12.9.

Statistical distribution theory also shows that an F random variable with 1 degree of freedom in the numerator is the square of a t random variable with the same degrees of freedom as the denominator of the F random variable. Thus, the F and t tests will always provide the same conclusions regarding the hypothesis test for a single independent variable in a multiple regression model.

Example 12.9 Housing Price Prediction for Small Cities (Hypothesis Tests for Coefficient Subsets)

The developers of the housing price prediction model from Example 12.8 wanted to determine if the combined effect of tax rate and percent commercial property contributes to the prediction after the effects of house size and income have been previously included. Data for this example are in the data file **Citydatr**.

Solution Continuing with the problem from Examples 12.7 and 12.8, we have a conditional test of the hypothesis that two variables are not significant predictors, given that the other two are significant predictors:

$$H_0 : \beta_3 = \beta_4 = 0 \mid \beta_1, \beta_2 \neq 0$$

This test will be conducted using the procedure in Equation 12.24. Figure 12.9 presents the regression for the complete model with all four predictor variables. In that regression $SSE = 1{,}149.14$. In Figure 12.11 we have the reduced regression with only house size and income as predictor variables. In that regression $SSE = 1{,}426.93$. The hypothesis is tested by first computing the F statistic whose numerator is the error sum of squares for the reduced model $[SSE(R)]$ minus the SSE for the complete model:

$$F = \frac{(1426.93 - 1149.14)/2}{13.52} = 10.27$$

The F statistic has 2 degrees of freedom—for the two variables being tested simultaneously—for the numerator and 85 degrees of freedom for the denominator. Note that the variance estimator, $s_e^2 = 13.52$, is obtained from the complete model in Figure 12.9, which has 85 degrees of freedom for error. The critical value for F with $\alpha = 0.01$ and 2 and 85 degrees of freedom, from Appendix Table 9, is approximately 4.9. Since the computed value of F exceeds the critical value, we reject the null hypothesis that tax rate and percent commercial property are not in combination conditionally significant. The combined effect of these two variables does improve the model that predicts housing price. Therefore, tax rate and percent commercial property should be included in the model.

Figure 12.11 Housing-Price Regression: Reduced Model (Minitab Output)

Regression Analysis: hseval versus sizehse, income72

```
The regression equation is
hseval = -42.2 + 9.14 sizehse + 0.00393 incom72



Predictor        Coef    SE Coef       T      P
Constant      -42.208      9.810   -4.30  0.000
sizehse         9.135      1.940    4.71  0.000
incom72      0.003927   0.001473    2.67  0.009


S = 4.04987    R-Sq = 34.7%   R-Sq(adj) = 33.2%

Analysis of Variance

Source            DF        SS       MS      F      P
Regression         2    759.70   379.85  23.16  0.000
Residual Error    87   1426.93    16.40
Total             89   2186.63

Source     DF    Seq SS                          SSE(R)
sizehse     1    643.12
incom72     1    116.58
```

We also computed this regression with the variable "comper" excluded and found that the resulting *SSE* was as follows:

$$SSE(1) = 1{,}185.29$$

Then the computed *F* statistic for this variable was as follows:

$$F = \frac{(1185.29 - 1149.14)/1}{13.52} = 2.674$$

The square root of 2.674 is 1.64, which is the computed *t* statistic for the variable Comper in the regression output in Figure 12.9. Using either the computed *F* or the computed *t*, we would obtain this result for the hypotheses for this variable:

$$H_0 : \beta_{Comper} = 0 \mid \beta_l \neq 0, l \neq Comper$$
$$H_1 : \beta_{Comper} \neq 0 \mid \beta_l \neq 0, l \neq Comper$$

Exercises

Basic Exercise

12.35 Suppose that you have estimated coefficients for the following regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Test the hypothesis that all three of the predictor variables are equal to 0, given the following analysis of variance tables:

a. Analysis of variance

Source	DF	SS	MS
Regression	3	4,500	
Residual error	26	500	

b. Analysis of variance

Source	DF	SS	MS
Regression	3	9,780	
Residual error	26	2,100	

c. Analysis of variance

Source	DF	SS	MS
Regression	3	46,000	
Residual error	26	25,000	

d. Analysis of variance

Source	DF	SS	MS
Regression	3	87,000	
Residual error	26	48,000	

Application Exercises

12.36 An aircraft company wanted to predict the number of worker-hours necessary to finish the design of a new plane. Relevant explanatory variables were thought to be the plane's top speed, its weight, and the number of parts it had in common with other models built by the company. A sample of 27 of the company's planes was taken, and the following model was estimated:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

y = design effort, in millions of worker-hours
x_1 = plane's top speed, in miles per hour
x_2 = plane's weight, in tons
x_3 = percentage of parts in common with other models

The estimated regression coefficients were as follows:

$$b_1 = 0.661 \quad b_2 = 0.065 \quad b_3 = -0.018$$

The total sum of squares and regression sum of squares were found to be as follows:

$$SST = 3.881 \quad \text{and} \quad SSR = 3.549$$

a. Test the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

b. Set out the analysis of variance table.

12.37 In a study of the influence of financial institutions on bond interest rates in Germany, quarterly data over a period of 12 years were analyzed. The postulated model was

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where

y = change over the quarter in the bond interest rates
x_1 = change over the quarter in bond purchases by financial institutions
x_2 = change over the quarter in bond sales by financial institutions

The estimated partial regression coefficients were as follows:

$$b_1 = 0.057 \quad b_2 = -0.065$$

The corrected coefficient of determination was found to be $R^2 = 0.463$. Test the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = 0$$

12.38 The following model was fitted to a sample of 30 families in order to explain household milk consumption:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where

y = milk consumption, in quarts per week
x_1 = weekly income, in hundreds of dollars
x_2 = family size

The least squares estimates of the regression parameters were as follows:

$$b_0 = -0.025 \quad b_1 = 0.052 \quad b_2 = 1.14$$

The estimated standard errors were as follows:

$$s_{b_1} = 0.023 \quad s_{b_2} = 0.35$$

The total sum of squares and regression sum of squares were found to be as follows:

$$SST = 162.1 \quad \text{and} \quad SSR = 88.2$$

a. Test the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = 0$$

b. Set out the analysis of variance table.

12.39 The following model was fitted to a sample of 25 students using data obtained at the end of their freshman year in college. The aim was to explain students' weight gains:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

y = weight gained, in pounds, during freshman year
x_1 = average number of meals eaten per week
x_2 = average number of hours of exercise per week
x_3 = average number of beers consumed per week

The least squares estimates of the regression parameters were as follows:

$$b_0 = 7.35 \quad b_1 = 0.653 \quad b_2 = -1.345 \quad b_3 = 0.613$$

The estimated standard errors were as follows:

$$s_{b_1} = 0.189 \quad s_{b_2} = 0.565 \quad s_{b_3} = 0.243$$

The regression sum of squares and error sum of squares were found to be as follows:

$$SSR = 79.2 \quad \text{and} \quad SSE = 45.9$$

a. Test the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

b. Set out the analysis of variance table.

12.40 A dependent variable is regressed on K independent variables, using n sets of sample observations. We denote SSE as the error sum of squares and R^2 as the coefficient of determination for this estimated regression. We want to test the null hypothesis that K_1 of these independent variables, taken together, do not linearly affect the dependent variable, given that the other $(K - K_1)$ independent variables are also to be used. Suppose that the regression is reestimated with the K_1 independent variables of interest excluded. Let SSE^* denote the error sum of squares and R^{*2}, the coefficient of determination for this regression. Show that the statistic for testing our null hypothesis, introduced in Section 12.5, can be expressed as follows:

$$\frac{(SSE^* - SSE)/K_1}{SSE/(n - K - 1)} = \frac{R^2 - R^{*2}}{1 - R^2} \cdot \frac{n - K - 1}{K_1}$$

12.41 The following model was fitted to a sample of 30 families in order to explain household milk consumption:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where

y = milk consumption, in quarts per week
x_1 = weekly income, in hundreds of dollars
x_2 = family size

The least squares estimates of the regression parameters were as follows:

$$b_0 = -0.025 \quad b_1 = 0.052 \quad b_2 = 1.14$$

The total sum of squares and regression sum of squares were found to be as follows:

$$SST = 162.1 \quad \text{and} \quad SSR = 88.2$$

A third independent variable—number of preschool children in the household—was added to the regression model. The sum of squared errors when this augmented model was estimated by least squares was found to be 83.7. Test the null hypothesis that, all other things being equal, the number of preschool children in the household does not linearly affect milk consumption.

12.42 Suppose that a dependent variable is related to K independent variables through a multiple regression model. Let R^2 denote the coefficient of determination and $\overline{R}^2$, the corrected coefficient. Suppose that n sets of observations are used to fit the regression.

a. Show that

$$\overline{R}^2 = \frac{(n-1)R^2 - K}{n-K-1}$$

b. Show that

$$R^2 = \frac{(n-K-1)\overline{R}^2 + K}{n-1}$$

c. Show that the statistic for testing the null hypothesis that all the regression coefficients are 0 can be written as

$$\frac{SSR/K}{SSE/(n-K-1)} = \frac{n-K-1}{K} \cdot \frac{\overline{R}^2 + A}{1-\overline{R}^2}$$

where

$$A = \frac{K}{n-K-1}$$

12.6 Prediction

An important application of regression models is to predict or forecast values of the dependent variable, given values for the independent variables. Forecasts can be computed directly from the estimated regression model using the coefficient estimates in that model, as shown in Equation 12.25.

Predictions from the Multiple Regression Models

Given that the population regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

holds and that the standard regression assumptions are valid, let $b_0, b_1, \ldots, b_K$ be the least squares estimates of the model coefficients, β_j, where $j = 1, \ldots, K$, based on the $x_1, x_2, \ldots, x_K$ $(i = 1, \ldots, n)$ data points. Then, given a new observation of a data point, $x_{1,n+1}, x_{2,n+1}, \ldots, x_{K,n+1}$ the best linear unbiased forecast of y_{n+1} is

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{1i} + \cdots + b_K x_{Ki} \quad i = n+1 \tag{12.25}$$

It is very risky to obtain forecasts that are based on X values outside the range of the data used to estimate the model coefficients because we do not have data evidence to support the linear model at those points.

In addition to the predicted value of Y for a particular set of x_j terms, we are often interested in a confidence interval or a prediction interval associated with the prediction. As we discussed in Section 11.6, the confidence interval includes the expected value of Y with probability $1 - \alpha$. In contrast, the prediction interval includes individual predicted values—expected values of Y plus the random error term. To obtain these intervals, we need to compute estimates of the standard deviations for the expected value of Y and for the individual points. These computations are similar in form to those used in simple regression, but the estimator equations are much more complicated. The standard deviations for predicted values, $s_{\hat{y}}$, are a function of the standard error of the estimate, s_e; the standard deviation of the predictor variables; the correlations between the predictor variables; and the square of the distance between the mean of the independent variables and the X terms for the prediction. This standard deviation is similar to the standard deviation for simple regression predictions in Chapter 11. However, the equations for multiple regression are very complex and are not presented here—instead, we compute the values using Minitab. The standard deviations for the prediction interval, the confidence interval, and the corresponding intervals are computed by most good statistics packages. Excel does not have the capability to compute the standard deviation of the predicted variables.

Example 12.10 Forecast of Savings and Loan Profit Margin (Regression Model Forecasts)

You have been asked to forecast the savings and loan profit margin for a year in which the percentage net revenue is 4.50 and there are 9,000 offices, using the savings and loan regression model. Data are stored in the file **Savings and Loan**.

Solution Using the notation from Equation 12.25, we have the following variables:

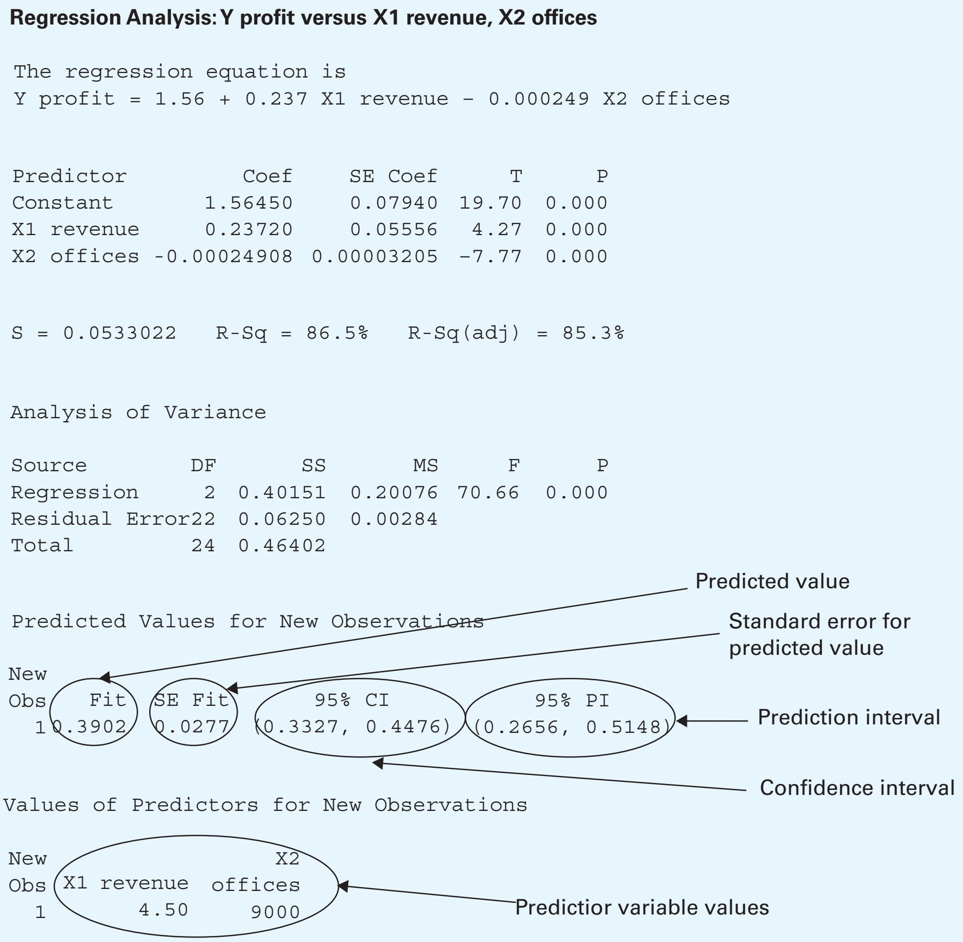
$$x_{1,n+1} = 4.50 \quad x_{2,n+1} = 9{,}000$$

Using these values, we find that our point predictor of profit margin is as follows:

$$\begin{aligned}\hat{y}_{n+1} &= b_0 + b_1 x_{1,n+1} + b_{2,n+1} \\ &= 1.565 + (0.237)(4.50) - (0.000249)(9{,}000) = 0.39\end{aligned}$$

Thus, for a year when the percentage net revenue per deposit dollar is 4.50 and the number of offices is 9,000, we predict that the profit margin for savings and loan associations will be 0.39.

Figure 12.12 Forecasts and Forecast Intervals for Multiple Regression (Minitab Output)



Regression Analysis: Y profit versus X1 revenue, X2 offices

```
The regression equation is
Y profit = 1.56 + 0.237 X1 revenue - 0.000249 X2 offices


Predictor          Coef     SE Coef      T      P
Constant        1.56450     0.07940  19.70  0.000
X1 revenue      0.23720     0.05556   4.27  0.000
X2 offices -0.00024908  0.00003205  -7.77  0.000


S = 0.0533022   R-Sq = 86.5%   R-Sq(adj) = 85.3%


Analysis of Variance

Source         DF       SS       MS      F      P
Regression      2  0.40151  0.20076  70.66  0.000
Residual Error22  0.06250  0.00284
Total          24  0.46402


Predicted Values for New Observations

New
Obs     Fit   SE Fit        95% CI               95% PI
  1  0.3902   0.0277  (0.3327, 0.4476)  (0.2656, 0.5148)


Values of Predictors for New Observations

New                      X2
Obs  X1 revenue     offices
  1        4.50        9000
```

Predicted value

Standard error for predicted value

Prediction interval

Confidence interval

Predictior variable values

Predicted values, confidence intervals, and prediction intervals can be computed directly in the Minitab regression routine.

The regression output is shown in Figure 12.12. The predicted value, $\hat{y} = 0.39$, and its standard deviation, 0.0277, are presented, along with the confidence interval and the prediction interval. The confidence interval—CI—provides an interval for the expected value of Y on the linear function defined by the values of the independent variables. This interval is a function of the standard error of the regression model, the distance that the x_j values are from their individual sample means, and the correlation between the x_j variables used to fit the model. The prediction interval—PI—provides an interval for a single observed value. Thus, it includes the variability associated with the expected value plus the variability of a single point about the predicted value.

Exercises

Visit **www.MyStatLab.com** or **www.pearsonhighered.com/newbold** to access the data files.

Basic Exercise

12.43 Given the estimated multiple regression equation

$$\hat{y} = 6 + 5x_1 + 4x_2 + 7x_3 + 8x_4$$

what is the predicted value of Y in each case?

a. $x_1 = 10, x_2 = 23, x_3 = 9$, and $x_4 = 12$
b. $x_1 = 23, x_2 = 18, x_3 = 10$, and $x_4 = 11$
c. $x_1 = 10, x_2 = 23, x_3 = 9$, and $x_4 = 12$
d. $x_1 = -10, x_2 = 13, x_3 = -8$, and $x_4 = -16$

Application Exercises

12.44 The following model was fitted to a sample of 25 students using data obtained at the end of their freshman year in college. The aim was to explain students' weight gains:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

y = weight gained, in pounds, during freshman year
x_1 = average number of meals eaten per week
x_2 = average number of hours of exercise per week
x_3 = average number of beers consumed per week

The least squares estimates of the regression parameters were as follows:

$$b_0 = 7.35 \quad b_1 = 0.653 \quad b_2 = -1.345 \quad b_3 = 0.613$$

Predict the weight gain for a freshman who eats an average of 20 meals per week, exercises an average of 10 hours per week, and consumes an average of 6 beers per week.

12.45 The following model was fitted to a sample of 30 families in order to explain household milk consumption:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where

y = milk consumption, in quarts per week
x_1 = weekly income, in hundreds of dollars
x_2 = family size

The least squares estimates of the regression parameters were as follows:

$$b_0 = -0.025 \quad b_1 = 0.052 \quad b_2 = 1.14$$

Predict the weekly milk consumption of a family of four with an income of \$600 per week.

12.46 An aircraft company wanted to predict the number of worker-hours necessary to finish the design of a new plane. Relevant explanatory variables were thought to be the plane's top speed, its weight, and the number of parts it had in common with other models built by the company. A sample of 27 of the company's planes was taken, and the following model was estimated:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

y = design effort, in millions of worker-hours
x_1 = plane's top speed, in miles per hour
x_2 = plane's weight, in tons
x_3 = percentage number of parts in common with other models

The estimated regression coefficients were as follows:

$$b_1 = 0.661 \quad b_2 = 0.065 \quad b_3 = -0.018$$

and the estimated intercept was 2.0.

Predict design effort for a plane with a top speed of Mach 1.0, weighing 7 tons, and having 50% of its parts in common with other models.

12.47 A real estate agent hypothesizes that in her town the selling price of a house in dollars (y) depends on its size in square feet of floor space (x_1), the lot size in square feet (x_2), the number of bedrooms (x_3), and the number of bathrooms (x_4). For a random sample of 20 house sales, the following least squares estimated model was obtained:

$$\hat{y} = 1998.5 + \underset{(2.5543)}{22.352x_1} + \underset{(1.4492)}{1.4686x_2} + \underset{(1820.8)}{6767.3x_3} + \underset{(1996.2)}{2701.1x_4}$$

$$R^2 = 0.9843$$

The numbers in parentheses under the coefficients are the estimated coefficient standard errors.

a. Interpret in the context of this model the estimated coefficient on x_2.
b. Interpret the coefficient of determination.
c. Assuming that the model is correctly specified, test, at the 5% level against the appropriate one-sided alternative, the null hypothesis that, all else being equal, selling price does not depend on number of bathrooms.
d. Estimate the selling price of a house with 1,250 square feet of floor space, a lot of 4,700 square feet, 3 bedrooms, and 1 bathroom.

12.48 Transportation Research, Inc., has asked you to prepare a multiple regression equation to estimate the effect of variables on fuel economy. The data for this study are contained in the data file **Motors**, and the dependent variable is miles per gallon—milpgal—as established by the Department of Transportation certification.

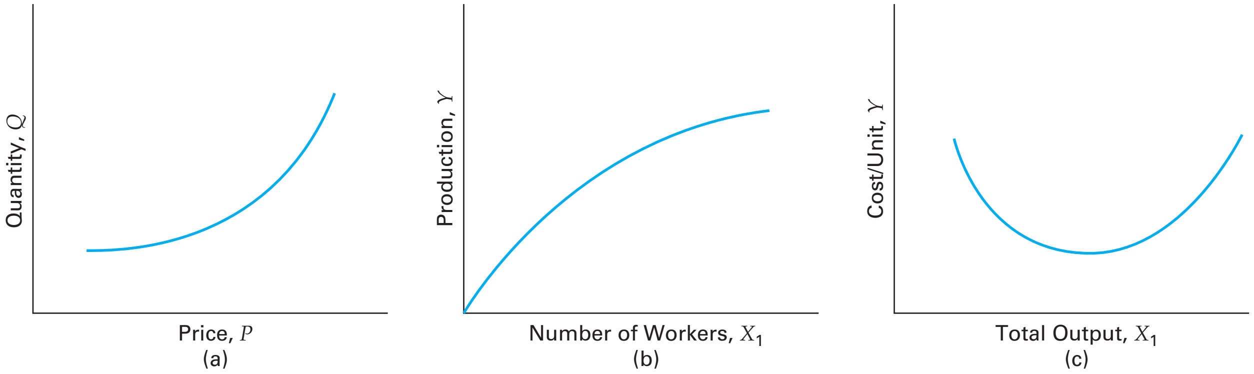
a. Prepare a regression equation that uses vehicle horsepower—horsepower—and vehicle weight—weight—as independent variables. Determine the predicted value, the confidence interval of the prediction, and the prediction interval when the horsepower is 140 and the vehicle weight is 3,000 pounds.
b. Prepare a second regression equation that adds the number of cylinders—cylinder—as an independent variable to the equation from part a. Determine the predicted value, the confidence interval of the prediction, and the prediction interval when the horsepower is 140, the number of cylinders is 6 and the vehicle weight is 3,000 pounds.

12.7 Transformations for Nonlinear Regression Models

We have seen how regression analysis can be used to estimate linear relationships that predict a dependent variable as a function of one or more independent variables. These applications are very important. However, in addition, there are a number of economic and business relationships that are not strictly linear. In this section we develop procedures for modifying certain nonlinear model formats so that multiple regression procedures can be used to estimate the model coefficients. Thus, our objective in Sections 12.7 and 12.8 is to expand the range of problems that are adaptable to regression analysis. In this way we see that regression analysis has even broader applications.

By examining the least squares algorithm, we will see that, with careful manipulation of nonlinear models, it is possible to use least squares for a broader set of applied problems. The assumptions concerning independent variables in multiple regression are not very restrictive. Independent variables define points at which we measure a random variable Y. We assume that there is a linear relationship between the levels of the independent variables X_j, where $j = 1, \dots, K$, and the expected value of the dependent variable Y. We can take advantage of this freedom to expand the set of models that can be estimated. Thus, we can move beyond linear models in our multiple regression applications. Three examples are shown in Figure 12.13:

Figure 12.13 Examples of Quadratic Functions



1. Supply functions may be nonlinear.
2. The increase in total output with increases in the number of workers may become flatter as more workers are added.
3. Average cost per unit produced is often minimized at an intermediate level of production.

Quadratic Transformations

We have spent considerable time developing regression analysis to estimate linear equations. There are also many processes that can best be represented by nonlinear equations. Total revenue has a quadratic relationship with price, with maximum revenue occurring at an intermediate price level if the demand function has a negative slope. In many cases the minimum production cost per unit occurs at an intermediate level of output, with cost per unit decreasing as we approach the minimum cost per unit and then increasing after passing the minimum unit cost level. We can model a number of these economic and business relationships by using a quadratic model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$$

To estimate the coefficients of a quadratic model for applications such as these, we can transform or modify the variables, as shown in Equations 12.26 and 12.27. In this way a nonlinear quadratic model is converted to a model that is linear in a modified set of variables.

Quadratic Model Transformations

The quadratic function

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon \qquad (12.26)$$

can be transformed into a linear multiple regression model by defining new variables:

$$z_1 = x_1$$
$$z_2 = x_1^2$$

and then specifying the model as

$$y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \varepsilon_i \qquad (12.27)$$

which is linear in the transformed variables. Transformed quadratic variables can be combined with other variables in a multiple regression model. Thus, we can fit a multiple quadratic regression using transformed variables. The goal is to find models that are linear in other mathematical forms of a variable.

By transforming the variables, we can estimate a linear multiple regression model and use the results as a nonlinear model. Inference procedures for transformed quadratic models are the same as those that we have previously developed for linear models. In this way we avoid confusion that would result if different statistical procedures were used for linear versus quadratic models. The coefficients must be combined for interpretation. Thus, if we have a quadratic model, then the effect of a variable, X, is indicated by the coefficients of both the linear and the quadratic terms. We can also perform a simple hypothesis test to determine if a quadratic model is an improvement over a linear model. The Z_2 or X_1^2 variable is merely an additional variable whose coefficient can be tested—$H_0: \beta_2 = 0$—using the conditional Student's t or F statistic. If a quadratic model fits the data better than a linear model, then the coefficient of the quadratic variable—$Z_2 = X_1^2$—will be significantly different from 0. The same approach applies if we have variables such as $Z_3 = X_1^3$ or $Z_4 = X_1^2 X_2$.

Example 12.11 Production Costs (Quadratic Model Estimation)

Arnold Sorenson, production manager of New Frontiers Instruments, Inc., was interested in estimating the mathematical relationship between the number of electronic assemblies produced during an 8-hour shift and the average cost per assembly. This function would then be used to estimate cost for various production order bids and to determine the production level that would minimize average cost. Data are found in the data file **Production Cost**.

Solution Arnold collected data from nine shifts during which the number of assemblies ranged from 100 to 900. In addition, he obtained the average cost per unit for those days from the accounting department. These data are presented in a scatter plot prepared using Excel, shown in Figure 12.14. As a result of his study of economics and his experience, Arnold suspected that the function might be quadratic with an intermediate minimum average cost. He designed his analysis to consider both a linear and a quadratic average production cost function.

Figure 12.14 Mean Production Cost as a Function of Number of Units

Number of Units	Mean Cost per Unit
100	5.11
210	4.42
290	4.07
415	3.52
509	3.33
613	3.44
697	3.77
806	4.07
908	4.28

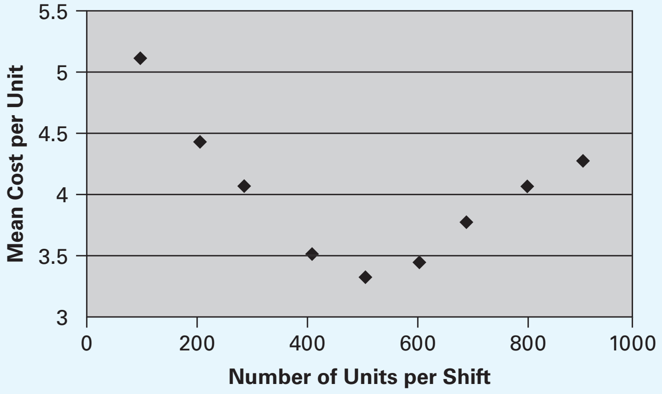


Figure 12.15 is the simple regression of cost as a linear function of the number of units. We see that the linear relationship is almost flat, indicating no linear relationship

Figure 12.15 Linear Regression Average Cost on Number of Units

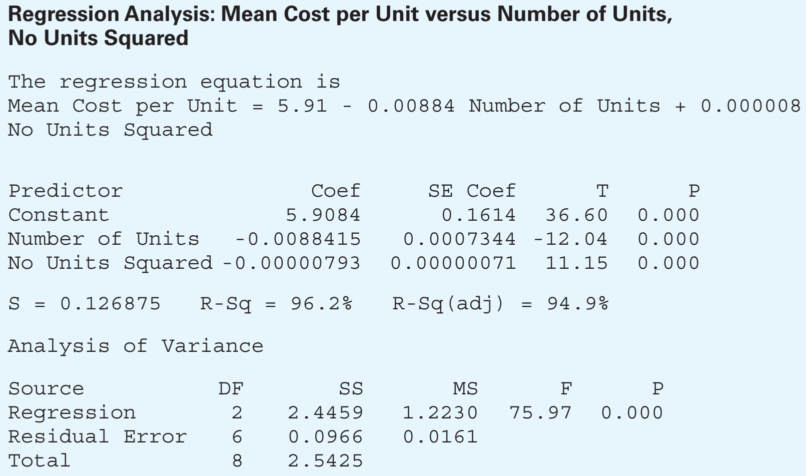
Regression Analysis: Mean Cost per Unit versus Number of Units

```
The regression equation is
Mean Cost per Unit = 4.43 - 0.000855 Number of Units


Predictor                Coef      SE Coef       T       P
Constant               4.4330       0.3994   11.10   0.000
Number of Units    -0.0008547    0.0007029   -1.22   0.263


S = 0.547614   R-Sq = 17.4%   R-Sq(adj) = 5.6%

Analysis of Variance

Source            DF        SS        MS       F       P
Regression         1    0.4433    0.4433    1.48   0.263
Residual Error     7    2.0992    0.2999
Total              8    2.5425
```

between average cost and number of units produced. If Arnold had simply used this relationship, he would have been led to serious errors in his cost-estimation procedures.

Figure 12.16 presents the quadratic regression that shows mean cost per unit as a nonlinear function of the number of units produced. Note that b_2 is different from 0 and thus should be included in the model. In addition, note that R^2 for the quadratic model is 0.962 compared to 0.174 for the linear model. By using the quadratic model, Arnold has produced a substantially more useful mean cost model.

Figure 12.16 Quadratic Model Analysis for Average Cost on Number of Units



Regression Analysis: Mean Cost per Unit versus Number of Units, No Units Squared

```
The regression equation is
Mean Cost per Unit = 5.91 - 0.00884 Number of Units + 0.000008
No Units Squared


Predictor                  Coef      SE Coef       T      P
Constant                 5.9084       0.1614   36.60  0.000
Number of Units      -0.0088415    0.0007344  -12.04  0.000
No Units Squared -0.00000793   0.00000071   11.15  0.000

S = 0.126875   R-Sq = 96.2%   R-Sq(adj) = 94.9%

Analysis of Variance

Source           DF        SS        MS       F       P
Regression        2    2.4459    1.2230   75.97   0.000
Residual Error    6    0.0966    0.0161
Total             8    2.5425
```

Logarithmic Transformations

A number of economic relationships can be modeled by exponential functions. For example, if the percent change in quantity of goods sold changes linearly in response to percent changes in the price, then the demand function will have an exponential form:

$$Q = \beta_0 P^{\beta_1}$$

where Q is the quantity demanded and P is the price per unit. Exponential demand functions have constant elasticity, and, thus, a 1% change in price results in the same percent change in quantity demanded for all price levels. In contrast, linear demand models indicate that a unit change in the price variable will result in the same change in quantity demanded for all price levels. Exponential demand models are widely used in the analysis of market behavior. One important feature of exponential models is that the coefficient β_1 is the constant elasticity, e, of demand Q with respect to price P:

$$e = \frac{\partial Q/Q}{\partial P/P} = \beta_1$$

This result is developed in most microeconomics textbooks. Exponential model coefficients are estimated using logarithmic transformations, as shown in Equation 12.29.

The logarithmic transformation assumes that the random error term multiplies the true value of Y to obtain the observed value. Thus, in the exponential model the error is a percentage of the true value, and the variance of the error distribution increases with increases in Y. If this result is not true, the log transformation is not correct. In that case a much more complex nonlinear estimation technique must be used. Those techniques are considerably beyond the scope of this book.