CHAPTER 4

New Text

MULTIPLE REGRESSION: ESTIMATION AND HYPOTHESIS TESTING

In the two-variable linear regression model that we have considered so far there was a single independent, or explanatory, variable. In this chapter we extend that model by considering the possibility that more than one explanatory variable may influence the dependent variable. A regression model with more than one explanatory variable is known as a **multiple regression model,** multiple because *multiple influences* (i.e., variables) may affect the dependent variable.

For example, consider the 1980s savings and loan (S&L) crisis resulting from the bankruptcies of some S&L institutions in several states. Similar events also occurred in the fall of 2008 as several banks were forced into bankruptcy. What factors should we focus on to understand these events? Is there a way to reduce the possibility that they will happen again? Suppose we want to develop a regression model to explain bankruptcy, the dependent variable. Now a phenomenon such as bankruptcy is too complex to be explained by a single explanatory variable; the explanation may entail several variables, such as the ratio of primary capital to total assets, the ratio of loans that are more than 90 days past due to total assets, the ratio of nonaccruing loans to total assets, the ratio of renegotiated loans to total assets, the ratio of net income to total assets, etc.[1] To include all these variables in a regression model to allow for the multiplicity of influences affecting bankruptcies, we have to consider a multiple regression model.

Needless to say, we could cite hundreds of examples of multiple regression models. In fact, most regression models are multiple regression models because very few economic phenomena can be explained by only a single explanatory variable, as in the case of the two-variable model.

[1]As a matter of fact, these were some of the variables that were considered by the Board of Governors of the Federal Reserve System in their internal studies of bankrupt banks.

In this chapter we discuss the multiple regression model seeking answers to the following questions:

1. How do we estimate the multiple regression model? Is the estimating procedure any different from that for the two-variable model?
2. Is the hypothesis-testing procedure any different from the two-variable model?
3. Are there any unique features of multiple regressions that we did not encounter in the two-variable case?
4. Since a multiple regression can have any number of explanatory variables, how do we decide how many variables to include in any given situation?

To answer these and other related questions, we first consider the simplest of the multiple regression models, namely, the three-variable model in which the behavior of the dependent variable Y is examined in relation to two explanatory variables, X_2 and X_3. Once the three-variable model is clearly understood, the extension to the four-, five-, or more variable case is quite straightforward, although the arithmetic gets a bit tedious. (But in this age of high-speed computers, that should not be a problem.) It is interesting that the three-variable model itself is in many ways a clear-cut extension of the two-variable model, as the following discussion reveals.

4.1 THE THREE-VARIABLE LINEAR REGRESSION MODEL

Generalizing the two-variable population regression function (PRF), we can write the three-variable PRF in its nonstochastic form as

$$E(Y_t) = B_1 + B_2X_{2t} + B_3X_{3t} \tag{4.1}$$[2]

and in the stochastic form as

$$Y_t = B_1 + B_2X_{2t} + B_3X_{3t} + u_t \tag{4.2}$$

$$= E(Y_t) + u_t \tag{4.3}$$

where Y = the dependent variable
X_2 and X_3 = the explanatory variables
u = the stochastic disturbance term
t = the tth observation

[2]Equation (4.1) can be written as: $E(Y_t) = B_1X_{1t} + B_2X_{2t} + B_3X_{3t}$ with the understanding that $X_{1t} = 1$ for each observation. The presentation in Eq. (4.1) is for notational convenience in that the subscripts on the parameters or their estimators match the subscripts on the variables to which they are attached.

In case the data are cross-sectional, the subscript i will denote the ith observation. Note that we introduce u in the three-variable, or, more generally, in the multivariable model for the same reason that it was introduced in the two-variable case.

B_1 is the intercept term. It represents the average value of Y when X_2 and X_3 are set equal to zero. The coefficients B_2 and B_3 are called **partial regression coefficients;** their meaning will be explained shortly.

Following the discussion in Chapter 2, Equation (4.1) gives the *conditional mean* value of Y, conditional upon the given or fixed values of the variables X_2 and X_3. Therefore, as in the two-variable case, multiple regression analysis is *conditional regression* analysis, conditional upon the given or fixed values of the explanatory variables, and we obtain *the average, or mean, value of Y for the fixed values of the X variables.* Recall that the PRF gives the (conditional) means of the Y populations corresponding to the given levels of the explanatory variables, X_2 and X_3.[3]

The stochastic version, Equation (4.2), states that any individual Y value can be expressed as the sum of two components:

1. A systematic, or *deterministic,* component ($B_1 + B_2X_{2t} + B_3X_{3t}$), which is simply its mean value $E(Y_t)$ (i.e., the point on the population regression line, PRL),[4] and
2. u_t, which is the *nonsystematic,* or *random,* component, determined by factors other than X_2 and X_3.

All this is familiar territory from the two-variable case; the only point to note is that we now have two explanatory variables instead of one explanatory variable.

Notice that Eq. (4.1), or its stochastic counterpart Eq. (4.2), is a *linear regression model*—a model that is *linear in the parameters,* the B's. As noted in Chapter 2, our concern in this book is with regression models that are linear in the parameters; such models may or may not be linear in the variables (but more on this in Chapter 5).

The Meaning of Partial Regression Coefficient

As mentioned earlier, the regression coefficients B_2 and B_3 are known as **partial regression** or **partial slope coefficients.** The meaning of the partial regression coefficient is as follows: B_2 measures the *change* in the mean value of Y, $E(Y)$, per unit change in X_2, holding the value of X_3 constant. Likewise, B_3 measures the

[3]Unlike the two-variable case, we cannot show this diagrammatically because to represent the three variables Y, X_2, and X_3, we have to use a three-dimensional diagram, which is difficult to visualize in two-dimensional form. But by stretching the imagination, we can visualize a diagram similar to Figure 2-6.

[4]Geometrically, the PRL in this case represents what is known as a plane.

change in the mean value of Y per unit change in X_3, holding the value of X_2 constant. This is the unique feature of a multiple regression; in the two-variable case, since there was only a single explanatory variable, we did not have to worry about the presence of other explanatory variables in the model. In the multiple regression model we want to find out what part of the change in the average value of Y can be directly attributable to X_2 and what part to X_3. Since this point is so crucial to understanding the logic of multiple regression, let us explain it by a simple example. Suppose we have the following PRF:

$$E(Y_t) = 15 - 1.2X_{2t} + 0.8X_{3t} \tag{4.4}$$

Let X_3 be held constant at the value 10. Putting this value in Equation (4.4), we obtain

$$\begin{aligned} E(Y_t) &= 15 - 1.2X_{2t} + 0.8(10) \\ &= (15 + 8) - 1.2X_{2t} \\ &= 23 - 1.2X_{2t} \end{aligned} \tag{4.5}$$

Here the slope coefficient $B_2 = -1.2$ indicates that the mean value of Y decreases by 1.2 per unit increase in X_2 when X_3 is held constant—in this example it is held constant at 10 although any other value will do.[5] This slope coefficient is called the *partial regression coefficient.*[6] Likewise, if we hold X_2 constant, say, at the value 5, we obtain

$$\begin{aligned} E(Y_t) &= 15 - 1.2(5) + 0.8X_{3t} \\ &= 9 + 0.8X_{3t} \end{aligned} \tag{4.6}$$

Here the slope coefficient $B_3 = 0.8$ means that the mean value of Y increases by 0.8 per unit increase in X_3 when X_2 is held constant—here it is held constant at 5, but any other value will do just as well. This slope coefficient too is a partial regression coefficient.

In short, then, a partial regression coefficient reflects the (partial) effect of one explanatory variable on the mean value of the dependent variable when the values of other explanatory variables included in the model are held constant. This unique feature of multiple regression enables us not only to include more than one explanatory variable in the model but also to "isolate" or "disentangle" the effect of each X variable on Y from the other X variables included in the model.

We will consider a concrete example in Section 4.5.

[5]As the algebra of Eq. (4.5) shows, it does not matter at what value X_3 is held constant, for that constant value multiplied by its coefficient will be a constant number, which will simply be added to the intercept.

[6]The mathematically inclined reader will notice at once that B_2 is the partial derivative of $E(Y)$ with respect to X_2 and that B_3 is the partial derivative of $E(Y)$ with respect to X_3.

4.2 ASSUMPTIONS OF THE MULTIPLE LINEAR REGRESSION MODEL

As in the two-variable case, our first order of business is to estimate the regression coefficients of the multiple regression model. Toward that end, we continue to operate within the framework of the classical linear regression model (CLRM) first introduced in Chapter 3 and to use the method of ordinary least squares (OLS) to estimate the coefficients.

Specifically, for model (4.2), we assume (cf. Section 3.1):

A4.1.

The regression model is linear in the parameters as in Eq. (4.1) and it is correctly specified.

A4.2.

X_2 and X_3 are uncorrelated with the disturbance term u. If X_2 and X_3 are nonstochastic (i.e., fixed numbers in repeated sampling), this assumption is automatically fulfilled.

However, if the X variables are random, or stochastic, they must be distributed independently of the error term u; otherwise, we will not be able to obtain unbiased estimates of the regression coefficients. But more on this in Chapter 11.

A4.3.

The error term u has a zero mean value; that is,

$$E(u_i) = 0 \tag{4.7}$$

A4.4.

Homoscedasticity, that is, the variance of u, is constant:

$$\text{var}(u_i) = \sigma^2 \tag{4.8}$$

A4.5.

No autocorrelation exists between the error terms u_i and u_j:

$$\text{cov}(u_i, u_j) \quad i \neq j \tag{4.9}$$

A4.6.

No exact collinearity exists between X_2 and X_3; that is, there is no exact linear relationship between the two explanatory variables. This is a new assumption and is explained later.

A4.7.

For hypothesis testing, the error term u follows the normal distribution with mean zero and (homoscedastic) variance σ^2. That is,

$$u_i \sim N(0, \sigma^2) \tag{4.10}$$

Except for Assumption (4.6), the rationale for the other assumptions is the same as that discussed for the two-variable linear regression. As noted in Chapter 3, we make these assumptions to facilitate the development of the subject. In Part II we will revisit these assumptions and see what happens if one or more of them are not fulfilled in actual applications.

According to Assumption (4.6) there is no exact linear relationship between the explanatory variables X_2 and X_3, technically known as the assumption of *no collinearity*, or no **multicollinearity,** if more than one exact linear relationship is involved. This concept is new and needs some explanation.

Informally, no perfect collinearity means that a variable, say, X_2, cannot be expressed as an exact linear function of another variable, say, X_3. Thus, if we can express

$$X_{2t} = 3 + 2X_{3t}$$

or

$$X_{2t} = 4X_{3t}$$

then the two variables are **collinear,** for there is an **exact linear relationship** between X_2 and X_3. Assumption (4.6) states that this should not be the case. The logic here is quite simple. If, for example, $X_2 = 4X_3$, then substituting this in Eq. (4.1), we see that

$$\begin{aligned} E(Y_t) &= B_1 + B_2(4X_{3t}) + B_3X_{3t} \\ &= B_1 + (4B_2 + B_3)X_{3t} \\ &= B_1 + AX_{3t} \end{aligned} \tag{4.11}$$

where

$$A = 4B_2 + B_3 \tag{4.12}$$

Equation (4.11) is a two-variable model, not a three-variable model. Now even if we can estimate Eq. (4.11) and obtain an estimate of A, there is no way that we can get individual estimates of B_2 or B_3 from the estimated A. Note that since Equation (4.12) is one equation with two unknowns we need two (independent) equations to obtain unique estimates of B_2 and B_3.

The upshot of this discussion is that in cases of perfect collinearity we cannot estimate the individual partial regression coefficients B_2 and B_3; in other words,

we cannot assess the individual effect of X_2 and X_3 on Y. But this is hardly surprising, for we really do not have two independent variables in the model.

Although, in practice, the case of perfect collinearity is rare, the cases of **high** or **near perfect collinearity** abound. In a later chapter (see Chapter 8) we will examine this case more fully. For now we merely require that two or more explanatory variables do not have exact linear relationships among them.

4.3 ESTIMATION OF THE PARAMETERS OF MULTIPLE REGRESSION

To estimate the parameters of Eq. (4.2), we use the ordinary least squares (OLS) method whose main features have already been discussed in Chapters 2 and 3.

Ordinary Least Squares Estimators

To find the OLS estimators, let us first write the sample regression function (SRF) corresponding to the PRF Eq. (4.2), as follows:

$$Y_t = b_1 + b_2X_{2t} + b_3X_{3t} + e_t \tag{4.13}$$

where, following the convention introduced in Chapter 2, e is the *residual term*, or simply the *residual*—the sample counterpart of u—and where the b's are the *estimators* of the population coefficients, the B's. More specifically,

$$\begin{aligned} b_1 &= \text{the estimator of } B_1 \\ b_2 &= \text{the estimator of } B_2 \\ b_3 &= \text{the estimator of } B_3 \end{aligned}$$

The sample counterpart of Eq. (4.1) is

$$\hat{Y}_t = b_1 + b_2X_{2t} + b_3X_{3t} \tag{4.14}$$

which is the *estimated* population regression line (PRL) (actually a plane).

As explained in Chapter 2, the OLS principle chooses the values of the unknown parameters in such a way that the residual sum of squares (RSS) $\sum e_t^2$ is as small as possible. To do this, we first write Equation (4.13) as

$$e_t = Y_t - b_1 - b_2X_{2t} - b_3X_{3t} \tag{4.15}$$

Squaring this equation on both sides and summing over the sample observations, we obtain

$$\text{RSS:} \quad \sum e_t^2 = \sum (Y_t - b_1 - b_2X_{2t} - b_3X_{3t})^2 \tag{4.16}$$

And in OLS we minimize this RSS (which is simply the sum of the squared difference between actual Y_t and estimated Y_t).

The minimization of Equation (4.16) involves the calculus technique of differentiation. Without going into detail, this process of differentiation gives us the following equations, known as (least squares) *normal equations,* to help estimate the unknowns[7] (compare them with the corresponding equations given for the two-variable case in Equations [2.14] and [2.15]):

$$\overline{Y} = b_1 + b_2\overline{X}_2 + b_3\overline{X}_3 \tag{4.17}$$

$$\sum YX_{2t} = b_1\sum X_{2t} + b_2\sum X_{2t}^2 + b_3\sum X_{2t}X_{3t} \tag{4.18}$$

$$\sum Y_tX_{3t} = b_1\sum X_{3t} + b_2\sum X_{2t}X_{3t} + b_3\sum X_{3t}^2 \tag{4.19}$$

where the summation is over the sample range 1 to n. Here we have three equations in three unknowns; the knowns are the variables Y and the X's and the unknowns are the b's. Ordinarily, we should be able to solve three equations with three unknowns. By simple algebraic manipulations of the preceding equations, we obtain the three OLS estimators as follows:

$$b_1 = \overline{Y} - b_2\overline{X}_2 - b_3\overline{X}_3 \tag{4.20}$$

$$b_2 = \frac{(\sum y_tx_{2t})\left(\sum x_{3t}^2\right) - (\sum y_tx_{3t})(\sum x_{2t}x_{3t})}{\left(\sum x_{2t}^2\right)\left(\sum x_{3t}^2\right) - (\sum x_{2t}x_{3t})^2} \tag{4.21}$$

$$b_3 = \frac{(\sum y_tx_{3t})\left(\sum x_{2t}^2\right) - (\sum y_tx_{2t})(\sum x_{2t}x_{3t})}{\left(\sum x_{2t}^2\right)\left(\sum x_{3t}^2\right) - (\sum x_{2t}x_{3t})^2} \tag{4.22}$$

where, as usual, lowercase letters denote deviations from sample mean values (e.g., $y_t = Y_t - \overline{Y}$).

You will notice the similarity between these equations and the corresponding ones for the two-variable case given in Eqs. (2.16) and (2.17). Also, notice the following features of the preceding equations: (1) Equations (4.21) and (4.22) are symmetrical in that one can be obtained from the other by interchanging the roles of x_2 and x_3, and (2) the denominators of these two equations are identical.

Variance and Standard Errors of OLS Estimators

Having obtained the OLS estimators of the intercept and partial regression coefficients, we can derive the variances and standard errors of these estimators in the manner of the two-variable model. These variances or standard errors give us some idea about the variability of the estimators from sample to sample. As in the two-variable case, we need the standard errors for two main

[7]The mathematical details can be found in Appendix 4A.1.

purposes: (1) to establish confidence intervals for the true parameter values and (2) to test statistical hypotheses. The relevant formulas, stated without proof, are as follows:

$$\text{var}\,(b_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3t}^2 + \bar{X}_3^2 \sum x_{2t}^2 - 2\bar{X}_2\bar{X}_3 \sum x_{2t}x_{3t}}{\sum x_{2t}^2 \sum x_{3t}^2 - (\sum x_{2t}x_{3t})^2}\right] \cdot \sigma^2 \tag{4.23}$$

$$\text{se}(b_1) = \sqrt{\text{var}\,(b_1)} \tag{4.24}$$

$$\text{var}\,(b_2) = \frac{\sum x_{3t}^2}{\left(\sum x_{2t}^2\right)\left(\sum x_{3t}^2\right) - (\sum x_{2t}\, x_{3t})^2} \cdot \sigma^2 \tag{4.25}$$

$$\text{se}(b_2) = \sqrt{\text{var}\,(b_2)} \tag{4.26}$$

$$\text{var}\,(b_3) = \frac{\sum x_{2t}^2}{\left(\sum x_{2t}^2\right)\left(\sum x_{3t}^2\right) - (\sum x_{2t}\, x_{3t})^2} \cdot \sigma^2 \tag{4.27}$$

$$\text{se}(b_3) = \sqrt{\text{var}\,(b_3)} \tag{4.28}$$

In all these formulas σ^2 is the (homoscedastic) variance of the population error term u_t. The OLS estimator of this unknown variance is

$$\hat{\sigma}^2 = \frac{\sum e_t^2}{n - 3} \tag{4.29}$$

This formula is a straightforward extension of its two-variable companion given in Equation (3.8) except that now the degrees of freedom (d.f.) are $(n - 3)$. This is because in estimating RSS, $\sum e_t^2$, we must first obtain b_1, b_2, and b_3, which consume 3 d.f. This argument is quite general. In the four-variable case the d.f. will be $(n - 4)$; in the five-variable case, $(n - 5)$; etc.

Also, note that the (positive) square root of $\hat{\sigma}^2$

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \tag{4.30}$$

is the *standard error of the estimate*, or the *standard error of the regression*, which, as noted in Chapter 3, is the standard deviation of Y values around the estimated regression line.

A word about computing $\sum e_t^2$. Since $\sum e_t^2 = \sum(Y_t - \hat{Y}_t)^2$, to compute this expression, one has first to compute $\hat{Y}_t$, which the computer does very easily. But there is a shortcut to computing the RSS (see Appendix 4A.2), which is

$$\sum e_t^2 = \sum y_t^2 - b_2 \sum y_t\, x_{2t} - b_3 \sum y_t\, x_{3t} \tag{4.31}$$

which can be readily computed once the partial slopes are estimated.

Properties of OLS Estimators of Multiple Regression

In the two-variable case we saw that under assumed conditions the OLS estimators are best linear unbiased estimators (BLUE). This property continues to hold for the multiple regression. Thus, each regression coefficient estimated by OLS is linear and unbiased—on the average it coincides with the true value. Among all such linear unbiased estimators, the OLS estimators have the least possible variance so that the true parameter can be estimated more accurately than by competing linear unbiased estimators. In short, the OLS estimators are efficient.

As the preceding development shows, in many ways the three-variable model is an extension of its two-variable counterpart, although the estimating formulas are a bit involved. These formulas get much more involved and cumbersome once we go beyond the three-variable model. In that case, we have to use matrix algebra, which expresses various estimating formulas more compactly. Of course, in this text matrix algebra is not used. Besides, today you rarely compute the estimates by hand; instead, you let the computer do the work.

4.4 GOODNESS OF FIT OF ESTIMATED MULTIPLE REGRESSION: MULTIPLE COEFFICIENT OF DETERMINATION, R^2

In the two-variable case we saw that r^2 as defined in Equation (3.38) measures the goodness of fit of the fitted sample regression line (SRL); that is, it gives the *proportion or percentage of the total variation in the dependent variable Y explained by the single explanatory variable X.* This concept of r^2 can be extended to regression models containing any number of explanatory variables. Thus, in the three-variable case we would like to know the proportion of the total variation in $Y(=\sum y_i^2)$ explained by X_2 and X_3 jointly. The quantity that gives this information is known as the **multiple coefficient of determination** and is denoted by the symbol $\boldsymbol{R^2}$**;** conceptually, it is akin to r^2.

As in the two-variable case, we have the identity (cf. Eq. 3.36):

$$\text{TSS} = \text{ESS} + \text{RSS} \tag{4.32}$$

where TSS = the total sum of squares of the dependent variable $Y(=\sum y_i^2)$
ESS = the explained sum of squares (i.e., explained by all the X variables)
RSS = the residual sum of squares

Also, as in the two-variable case, R^2 is defined as

$$R^2 = \frac{\text{ESS}}{\text{TSS}} \tag{4.33}$$

That is, it is the ratio of the explained sum of squares to the total sum of squares; the only change is that the ESS is now due to more than one explanatory variable.

Now it can be shown that[8]

$$\text{ESS} = b_2 \sum y_t x_{2t} + b_3 \sum y_t x_{3t} \tag{4.34}$$

and, as shown before,

$$\text{RSS} = \sum y_t^2 - b_2 \sum y_t x_{2t} - b_3 \sum y_t x_{3t} \tag{4.35}$$

Therefore, R^2 can be computed as

$$R^2 = \frac{b_2 \sum y_t x_{2t} + b_3 \sum y_t x_{3t}}{\sum y_t^2} \tag{4.36}$$[9]

In passing, note that the positive square root of R^2, **R**, is known as the **coefficient of multiple correlation**, the two-variable analogue of r. Just as r measures the degree of linear association between Y and X, R can be interpreted as the degree of linear association between Y and all the X variables jointly. Although r can be positive or negative, R is always taken to be positive. In practice, however, R is of little importance.

4.5 ANTIQUE CLOCK AUCTION PRICES REVISITED

Let us take time out to illustrate all the preceding theory with the antique clock auction prices example we considered in Chapter 2 (See Table 2-14). Let Y = auction price, X_2 = age of clock, and X_3 = number of bidders. A priori, one would expect a positive relationship between Y and the two explanatory variables. The results of regressing Y on the two explanatory variables are as follows (the EViews output of this regression is given in Appendix 4A.4).

$$\begin{aligned}
\hat{Y}_i &= -1336.049 + 12.7413 X_{2i} + 85.7640 X_{3i} \\
\text{se} &= (175.2725) \quad (0.9123) \quad (8.8019) \\
t &= (-7.6226) \quad (13.9653) \quad (9.7437) \\
p &= (0.0000)^* \quad (0.0000)^* \quad (0.0000)^* \\
R^2 &= 0.8906; \quad F = 118.0585
\end{aligned} \tag{4.37}$$

Interpretation of the Regression Results

As expected, the auction price is positively related to both the age of the clock and the number of bidders. The interpretation of the slope coefficient of about 12.74 means that holding other variables constant, if the age of the clock goes up

[8] See Appendix 4A.2.

[9] R^2 can also be computed as $1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum e_i^2}{\sum y_i^2}$.

*Denotes an extremely small value.

by a year, the average price of the clock will go up by about 12.74 marks. Likewise, holding other variables constant, if the number of bidders increases by one, the average price of the clock goes up by about 85.76 marks. The negative value of the intercept has no viable economic meaning. The R^2 value of about 0.89 means that the two explanatory variables account for about 89 percent of the variation in the auction bid price, a fairly high value. The F value given in Eq. (4.37) will be explained shortly.

4.6 HYPOTHESIS TESTING IN A MULTIPLE REGRESSION: GENERAL COMMENTS

Although R^2 gives us an overall measure of goodness of fit of the estimated regression line, by itself R^2 does not tell us whether the estimated partial regression coefficients are statistically significant, that is, statistically different from zero. Some of them may be and some may not be. How do we find out?

To be specific, let us suppose we want to entertain the hypothesis that age of the antique clock has no effect on its price. In other words, we want to test the null hypothesis: H_0: $B_2 = 0$. How do we go about it? From our discussion of hypothesis testing for the two-variable model given in Chapter 3, in order to answer this question we need to find out the sampling distribution of b_2, the estimator of B_2. What is the sampling distribution of b_2? And what is the sampling distribution of b_1 and b_3?

In the two-variable case we saw that the OLS estimators, b_1 and b_2, are normally distributed if we are willing to assume that the error term u follows the normal distribution. Now in Assumption (4.7) we have stated that even for multiple regression we will continue to assume that u is normally distributed with zero mean and constant variance σ^2. Given this and the other assumptions listed in Section 4.2, we can prove that b_1, b_2, and b_3 *each* follow the normal distribution with means equal to B_1, B_2, and B_3, respectively, and the variances given by Eqs. (4.23), (4.25), and (4.27), respectively.

However, as in the two-variable case, if we replace the true but unobservable σ^2 by its unbiased estimator $\hat{\sigma}^2$ given in Eq. (4.29), the OLS estimators follow the t distribution with $(n - 3)$ d.f., not the normal distribution. That is,

$$t = \frac{b_1 - B_1}{\text{se}(b_1)} \sim t_{n-3} \tag{4.38}$$

$$t = \frac{b_2 - B_2}{\text{se}(b_2)} \sim t_{n-3} \tag{4.39}$$

$$t = \frac{b_3 - B_3}{\text{se}(b_2)} \sim t_{n-3} \tag{4.40}$$

Notice that the d.f. are now $(n - 3)$ because in computing the RSS, $\sum e_i^2$, and hence $\hat{\sigma}^2$, we first need to estimate the intercept and the two partial slope coefficients; so we lose 3 d.f.

We know that by replacing σ^2 with $\hat{\sigma}^2$ the OLS estimators follow the t distribution. Now we can use this information to establish confidence intervals as well as to test statistical hypotheses about the true partial regression coefficients. The actual mechanics in many ways resemble the two-variable case, which we now illustrate with an example.

4.7 TESTING HYPOTHESES ABOUT INDIVIDUAL PARTIAL REGRESSION COEFFICIENTS

Suppose in our illustrative example we hypothesize that

$$H_0: B_2 = 0 \quad \text{and} \quad H_1: B_2 \neq 0$$

That is, under the null hypothesis, the age of the antique clock has no effect whatsoever on its bid price, whereas under the alternative hypothesis, it is contended that age has some effect, positive or negative, on price. The alternative hypothesis is thus two-sided.

Given the preceding null hypothesis, we know that

$$\begin{aligned} t &= \frac{b_2 - B_2}{\text{se}(b_2)} \\ &= \frac{b_2}{\text{se}(b_2)} \qquad (\textit{Note: } B_2 = 0) \end{aligned} \tag{4.41}$$

follows the t distribution with $(n - 3) = 29$ d.f., since $n = 32$ in our example. From the regression results given in Eq. (4.37), we obtain

$$t = \frac{12.7413}{0.9123} \approx 13.9653 \tag{4.42}$$

which has the t distribution with 29 d.f.

On the basis of the computed t value, do we reject the null hypothesis that the age of the antique clock has no effect on its bid price? To answer this question, we can either use the *test of significance approach* or the *confidence interval approach,* as we did for the two-variable regression.

The Test of Significance Approach

Recall that in the test of significance approach to hypothesis testing we develop a test statistic, find out its sampling distribution, choose a level of significance α, and determine the critical value(s) of the test statistic at the chosen level of significance. Then we compare the value of the test statistic obtained from the sample at hand with the critical value(s) and reject the null hypothesis if the computed value of the test statistic exceeds the critical value(s).[10] Alternatively,

[10]If the test statistic has a negative value, we consider its absolute value and say that if the absolute value of the test statistic exceeds the critical value, we reject the null hypothesis.

we can find the p value of the test statistic and reject the null hypothesis if the p value is smaller than the chosen α value. The approach that we followed for the two-variable case also carries over to the multiple regression.

Returning to our illustrative example, we know that the test statistic in question is the t statistic, which follows the t distribution with $(n - 3)$ d.f. Therefore, we use the *t test of significance*. The actual mechanics are now straightforward. Suppose we choose $\alpha = 0.05$ or 5%. Since the alternative hypothesis is two-sided, we have to find the critical t value at $\alpha/2 = 2.5\%$ (Why?) for $(n - 3)$ d.f., which in the present example is 29. Then from the t table we observe that for 29 d.f.,

$$(-2.045 \leq t \leq 2.045) = 0.95 \tag{4.43}$$

That is, the probability that a t value lies between the limits -2.045 and $+2.045$ (i.e., the critical t values) is 95 percent.

From Eq. (4.42), we see that the computed t value under $H_0 : B_2 = 0$ is approximately 14, which obviously exceeds the critical t value of 2.045. We therefore reject the null hypothesis and conclude that age of an antique clock definitely has an influence on its bid price. This conclusion is also reinforced by the p value given in Eq. (4.37), which is practically zero. That is, if the null hypothesis that $B_2 = 0$ were true, our chances of obtaining a t value of about 14 or greater would be practically nil. Therefore, we can reject the null hypothesis more resoundingly on the basis of the p value than the conventionally chosen α value of 1% or 5%.

One-Tail or Two-Tail *t* Test? Since, a priori, we expect the coefficient of the age variable to be positive, we should in fact use the one-tail t test here. The 5% critical t value for the one-tail test for 29 d.f. now becomes 1.699. Since the computed t value of about 14 is still so much greater than 1.699, we reject the null hypothesis and now conclude that the age of the antique clock *positively* impacts its bid price; the two-tail test, on the other hand, simply told us that age of the antique clock could have a positive or negative impact on its bid price. Therefore, be careful about how you formulate your null and alternative hypotheses. Let theory be the guide in choosing these hypotheses.

The Confidence Interval Approach to Hypothesis Testing

The basics of the confidence interval approach to hypothesis testing have already been discussed in Chapter 3. Here we merely illustrate it with our numerical example. We showed previously that

$$P(-2.045 \leq t \leq 2.045) = 0.95$$

We also know from Eq. (4.39) that

$$t = \frac{b_2 - B_2}{\text{se}(b_2)}$$

If we substitute this t value into Equation (4.43), we obtain

$$P\left(-2.045 \leq \frac{b_2 - B_2}{\text{se}(b_2)} \leq 2.045\right) = 0.95$$

Which, after rearranging becomes

$$P[b_2 - 2.045\,\text{se}(b_2) \leq B_2 \leq b_2 + 2.045\,\text{se}(b_2)] = 0.95 \tag{4.44}$$

which is a 95% confidence interval for B_2 (cf. Eq. [3.26]). Recall that under the confidence interval approach, if the confidence interval, which we call the *acceptance region*, includes the null-hypothesized value, we do not reject the null hypothesis. On the other hand, if the null-hypothesized value lies outside the confidence interval, that is, in the *region of rejection*, we can reject the null hypothesis. But always bear in mind that in making either decision we are taking a chance of being wrong α% (say, 5%) of the time.

For our illustrative example, Eq. (4.44) becomes

$$12.7413 - 2.045(0.9123) \leq B_2 \leq 12.7413 + 2.045(0.9123)$$

that is,

$$10.8757 \leq B_2 \leq 14.6069 \tag{4.45}$$

which is a 95% confidence interval for true B_2. Since this interval does not include the null-hypothesized value, we can reject the null hypothesis: If we construct confidence intervals like expression (4.45), then 95 out of 100 such intervals will include the true B_2, but, as noted in Chapter 3, *we cannot say that the probability is 95% that the particular interval in Eq. (4.45) does or does not include the true* B_2.

Needless to say, we can use the two approaches to hypothesis testing to test hypotheses about any other coefficient given in the regression results for our illustrative example. As you can see from the regression results, the variable, number of bidders, is also statistically significant (i.e., significantly different from zero) because the estimated t value of about 8 has a p value of almost zero. Remember that the lower the p value, the greater the evidence against the null hypothesis.

4.8 TESTING THE JOINT HYPOTHESIS THAT $B_2 = B_3 = 0$ OR $R^2 = 0$

For our illustrative example we saw that *individually* the partial slope coefficients b_2 and b_3 are statistically significant; that is, *individually* each partial slope coefficient is significantly different from zero. But now consider the following null hypothesis:

$$H_0: B_2 = B_3 = 0 \tag{4.46}$$

This null hypothesis is a **joint hypothesis** that B_2 and B_3 are *jointly* or *simultaneously* (and not individually or singly) equal to zero. This hypothesis states that the two explanatory variables *together* have no influence on Y. This is the same as saying that

$$H_0: R^2 = 0 \tag{4.47}$$

That is, the two explanatory variables explain zero percent of the variation in the dependent variable (recall the definition of R^2). Therefore, the two sets of hypotheses (4.46) and (4.47) are equivalent; one implies the other. A test of either hypothesis is called a **test of the overall significance of the estimated multiple regression;** that is, whether Y is linearly related to both X_2 and X_3.

How do we test, say, the hypothesis given in Equation (4.46)? The temptation here is to state that since individually b_2 and b_3 are statistically different from zero in the present example, then jointly or collectively they also must be statistically different from zero; that is, we reject H_0 given in Eq. (4.46). In other words, since age of the antique clock and the number of bidders at the auction *each* has a significant effect on the auction price, *together* they also must have a significant effect on the auction price. But we should be careful here for, as we show more fully in Chapter 8 on multicollinearity, in practice, in a multiple regression one or more variables *individually* have no effect on the dependent variable but *collectively* they have a significant impact on it. *This means that the t-testing procedure discussed previously, although valid for testing the statistical significance of an individual regression coefficient, is not valid for testing the joint hypothesis.*

How then do we test a hypothesis like Eq. (4.46)? This can be done by using a technique known as **analysis of variance (ANOVA).** To see how this technique is employed, recall the following identity:

$$\text{TSS} = \text{ESS} + \text{RSS} \tag{4.32}$$

That is,

$$\sum y_t^2 = b_2 \sum y_t x_{2t} + b_3 \sum y_t x_{3t} + \sum e_t^2 \tag{4.48}$$[11]

Equation (4.48) *decomposes* the TSS into two components, one explained by the (chosen) regression model (ESS) and the other not explained by the model (RSS). *A study of these components of TSS is known as the analysis of variance (ANOVA) from the regression viewpoint.*

As noted in Appendix C every sum of squares has associated with it its degrees of freedom (d.f.); that is, the number of independent observations on

[11]This is Equation (4.35) written differently.

TABLE 4-1 ANOVA TABLE FOR THE THREE-VARIABLE REGRESSION

Source of variation	Sum of squares (SS)	d.f.	MSS $= \frac{\text{SS}}{\text{d.f.}}$
Due to regression (ESS)	$b_2 \sum y_t x_{2t} + b_3 \sum y_t x_{3t}$	2	$\frac{b_2 \sum y_t x_{2t} + b_3 \sum y_t x_{3t}}{2}$
Due to residual (RSS)	$\sum e_t^2$	$n - 3$	$\frac{\sum e_t^2}{n-3}$
Total (TSS)	$\sum y_t^2$	$n - 1$	

Note: MSS = mean, or average, sum of squares.

the basis of which the sum of squares is computed. Now each of the preceding sums of squares has these d.f.:

Sum of squares	d.f.
TSS	$n - 1$ (always, Why?)
RSS	$n - 3$ (three-variable model)
ESS	2 (three-variable model)*

*An easy way to find the d.f. for ESS is to subtract the d.f. for RSS from the d.f. for TSS.

Let us arrange all these sums of squares and their associated d.f. in a tabular form, known as the ANOVA table, as shown in Table 4-1.

Now given the assumptions of the CLRM (and Assumption 4.7) and the null hypothesis: $H_0: B_2 = B_3 = 0$, it can be shown that the variable

$$\begin{aligned} F &= \frac{\text{ESS/d.f.}}{\text{RSS/d.f.}} \\ &= \frac{\text{variance explained by } X_2 \text{ and } X_3}{\text{unexplained variance}} \\ &= \frac{(b_2 \sum y_t x_{2t} + b_3 \sum y_t x_{3t})/2}{\sum e_t^2/(n-3)} \end{aligned} \tag{4.49}$$

follows the F distribution with 2 and $(n - 3)$ d.f. in the numerator and denominator, respectively. (See Appendix C for a general discussion of the F distribution and Appendix D for some applications). *In general, if the regression model has k explanatory variables including the intercept term, the F ratio has (k − 1) d.f. in the numerator and (n − k) d.f. in the denominator.*[12]

How can we use the F ratio of Equation (4.49) to test the joint hypothesis that both X_2 and X_3 have no impact on Y? The answer is evident in Eq. (4.49). If the

[12]A simple way to remember this is that the numerator d.f. of the F ratio is equal to the number of *partial slope coefficients* in the model, and the denominator d.f. is equal to n minus the total number of parameters estimated (i.e., partial slopes plus the intercept).

TABLE 4-2 ANOVA TABLE FOR THE CLOCK AUCTION PRICE EXAMPLE

Source of variation	Sum of squares (SS)	d.f.	MSS $= \frac{SS}{d.f.}$
Due to regression (ESS)	4278295.3	2	4278295.3/2
Due to residual (RSS)	525462.2	29	525462.2/29
Total (TSS)	4803757.5	31	

$F = 2139147.6/18119.386 = 118.0585$*

*Figures have been rounded.

numerator of Eq. (4.49) is larger than its denominator—if the variance of Y explained by the regression (i.e., by X_2 and X_3) is larger than the variance not explained by the regression—the F value will be greater than 1. Therefore, as the variance explained by the X variables becomes increasingly larger relative to the unexplained variance, the F ratio will be increasingly larger, too. *Thus, an increasingly large F value will be evidence against the null hypothesis that the two (or more) explanatory variables have no effect on Y.*

Of course, this intuitive reasoning can be formalized in the usual framework of hypothesis testing. As shown in Appendix C, Section C.4, we compute F as given in Eq. (4.49) and compare it with the critical F value for 2 and $(n - 3)$ d.f. at the chosen level of α, the probability of committing a type I error. *As usual, if the computed F value exceeds the critical F value, we reject the null hypothesis that the impact of all explanatory variables is simultaneously equal to zero. If it does not exceed the critical F value, we do not reject the null hypothesis that the explanatory variables have no impact whatsoever on the dependent variable.*

To illustrate the actual mechanics, let us return to our illustrative example. The numerical counterpart of Table 4-1 is given in Table 4-2.

The entries in this table are obtained from the EViews computer output given in Appendix 4A.4.[13] From this table and the computer output, we see that the estimated F value is 118.0585, or about 119. Under the null hypothesis that $B_2 = B_3 = 0$, and given the assumptions of the classical linear regression model (CLRM), we know that the computed F value follows the F distribution with 2 and 29 d.f. in the numerator and denominator, respectively. If the null hypothesis were true, what would be the probability of our obtaining an F value of as much as 118 or greater for 2 and 13 d.f.? The p value of obtaining an F value of 118 or greater is 0.000000, which is practically zero. Hence, we can reject the null hypothesis that age and number of bidders *together* has no effect on the bid price of antique clocks.[14]

In our illustrative example it so happens that not only do we reject the null hypothesis that B_2 and B_3 are *individually* statistically insignificant, but we also

[13]Unlike other software packages, EViews does not produce the ANOVA table, although it gives the F value. But it is very easy to construct this table, for EViews gives TSS and RSS from which ESS can be easily obtained.

[14]If you had chosen $\alpha = 1\%$, the critical F value for 2 and 30 (which is close to 29) d.f. would be 5.39. The F value of 118 is obviously much greater than this critical value.

reject the hypothesis that *collectively* they are insignificant. However, this need not happen all the time. We will come across cases where not all explanatory variables individually have much impact on the dependent variable (i.e., some of the t values may be statistically insignificant) yet all of them *collectively* influence the dependent variable (i.e., the F test will reject the null hypothesis that all partial slope coefficients are simultaneously equal to zero.) As we will see, this happens if we have the problem of multicollinearity, which we will discuss more in Chapter 8.

An Important Relationship between F and R^2

There is an important relationship between the coefficient of determination R^2 and the F ratio used in ANOVA. This relationship is as follows:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \tag{4.50}$$

where n = the number of observations and k = the number of explanatory variables including the intercept.

Equation (4.50) shows how F and R^2 are related. These two statistics vary directly. When $R^2 = 0$ (i.e., no relationship between Y and the X variables), F is zero ipso facto. The larger R^2 is, the greater the F value will be. In the limit when $R^2 = 1$, the F value is infinite.

Thus the ***F* test** discussed earlier, which is a measure of the overall significance of the estimated regression line, is also a test of significance of R^2; that is, whether R^2 is different from zero. In other words, testing the null hypothesis Eq. (4.46) is equivalent to testing the null hypothesis that (the population) R^2 is zero, as noted in Eq. (4.47).

One advantage of the F test expressed in terms of R^2 is the ease of computation. All we need to know is the R^2 value, which is routinely computed by most regression programs. Therefore, the overall F test of significance given in Eq. (4.49) can be recast in terms of R^2 as shown in Eq. (4.50), and the ANOVA Table 4-1 can be equivalently expressed as Table 4-3.

TABLE 4-3 ANOVA TABLE IN TERMS OF R^2

Source of variation	Sum of squares (SS)	d.f.	MSS $= \frac{\text{SS}}{\text{d.f.}}$
Due to regression (ESS)	$R^2\left(\sum y_i^2\right)$	2	$\frac{R^2\left(\sum y_i^2\right)}{2}$
Due to residual (RSS)	$(1-R^2)\left(\sum y_i^2\right)$	$n-3$	$\frac{(1-R^2)\left(\sum y_i^2\right)}{(n-3)}$
Total (TSS)	$\sum y_i^2$	$n-1$	

Note: In computing the F value, we do not need to multiply R^2 and $(1-R^2)$ by $\sum y_i^2$ since it drops out, as can be seen from Eq. (4.49).

In the k-variable model the d.f. will be $(k-1)$ and $(n-k)$, respectively.

For our illustrative example, $R^2 = 0.8906$. Therefore, the F ratio of Equation (4.50) becomes

$$F = \frac{0.8906/2}{(1 - 0.8906)/29} \approx 118.12 \tag{4.51}$$

which is about the same F as shown in Table 4-2, except for rounding errors.

It is left for you to set up the ANOVA table for our illustrative example in the manner of Table 4-3.

4.9 TWO-VARIABLE REGRESSION IN THE CONTEXT OF MULTIPLE REGRESSION: INTRODUCTION TO SPECIFICATION BIAS

Let us return to our example. In Example 2.5, we regressed auction price on the age of the antique clock and the number of bidders separately, as shown in Equations (2.27) and (2.28). These equations are reproduced here with the usual regression output.

$$\begin{aligned} \hat{Y}_i &= -191.6662 + 10.4856\ Age_i \\ \text{se} &= (264.4393) + (1.7937) \\ t &= (-0.7248) \quad (5.8457) \qquad r^2 = 0.5325;\ F = 34.1723 \end{aligned} \tag{4.52}$$

$$\begin{aligned} \hat{Y}_i &= 807.9501 + 54.5724\ Bidders \\ \text{se} &= (231.9501) \quad (23.5724) \\ t &= (3.4962) \quad (2.3455) \qquad r^2 = 0.1549;\ F = 5.5017 \end{aligned} \tag{4.53}$$

If we compare these regressions with the results of the multiple regression given in Eq. (4.37), we see several differences:

1. The slope values in Equations (4.52) and (4.53) are different from those given in the multiple regression (4.37), especially that of the number of bidders variable.
2. The intercept values in the three regressions are also different.
3. The R^2 value in the multiple regression is quite different from the r^2 values given in the two bivariate regressions. In a bivariate regression, however, R^2 and r^2 are basically indistinguishable.

As we will show, some of these differences are statistically significant and some others may not be.

Why the differences in the results of the two regressions? Remember that in Eq. (4.37), while deriving the impact of age of the antique clock on the auction price, we held the number of bidders constant, whereas in Eq. (4.52) we simply neglected the number of bidders. Put differently, in Eq. (4.37) the effect of a clock's age on auction price is *net* of the effect, or influence, of the number of bidders, whereas in Eq. (4.52) the effect of the number of bidders has *not* been netted out. Thus, the coefficient of the age variable in Eq. (4.52) reflects the *gross*

effect—the *direct effect* of age as well as the *indirect effect* of the number of bidders. This difference between the results of regressions (4.37) and (4.52) shows very nicely the meaning of the "partial" regression coefficient.

We saw in our discussion of regression (4.37) that both the age of the clock and the number of bidders variables were *individually* as well as *collectively* important influences on the auction price. Therefore, by omitting the number of bidders variable from regression (4.52) we have committed what is known as a **(model) specification bias** or **specification error,** more specifically, the specification error of omitting a relevant variable from the model. Similarly, by omitting the age of the clock from regression (4.53), we also have committed a specification error.

Although we will examine the topic of specification errors in Chapter 7, what is important to note here is that you should be very careful in developing a regression model for empirical purposes. Take whatever help you can from the underlying theory and/or prior empirical work in developing the model. And once you choose a model, do not drop variables from the model arbitrarily.

4.10 COMPARING TWO R^2 VALUES: THE ADJUSTED R^2

By examining the R^2 values of our two-variable (Eq. [4.52] or Eq. [4.53]) and three-variable (Eq. [4.37]) regressions for our illustrative example, you will notice that the R^2 value of the former (0.5325 for Eq. [4.52] or 0.1549 for Eq. [4.53]) is smaller than that of the latter (0.8906). Is this always the case? Yes! An important property of R^2 is that the larger the number of explanatory variables in a model, the higher the R^2 will be. It would then seem that if we want to explain a substantial amount of the variation in a dependent variable, we merely have to go on adding more explanatory variables!

However, do not take this "advice" too seriously because the definition of $R^2 = \text{ESS/TSS}$ does not take into account the d.f. Note that in a k-variable model including the intercept term the d.f. for ESS is $(k - 1)$. Thus, if you have a model with 5 explanatory variables including the intercept, the d.f. associated with ESS will be 4, whereas if you had a model with 10 explanatory variables including the intercept, the d.f. for the ESS would be 9. But the conventional R^2 formula does not take into account the differing d.f. in the various models. Note that the d.f. for TSS is always $(n - 1)$. (Why?) *Therefore, comparing the R^2 values of two models with the same dependent variable but with differing numbers of explanatory variables is essentially like comparing apples and oranges.*

Thus, what we need is a measure of goodness of fit that is adjusted for (i.e., takes into account explicitly) the number of explanatory variables in the model. Such a measure has been devised and is known as the **adjusted R^2,** denoted by the symbol, $\overline{R}^2$. This $\overline{R}^2$ can be derived from the conventional R^2 (see Appendix 4A.3) as follows:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n - 1}{n - k} \tag{4.54}$$

Note that the R^2 we have considered previously is also known as the *unadjusted* R^2 for obvious reasons.

The features of the adjusted R^2 are:

1. If $k > 1$, $\overline{R}^2 \leq R^2$; that is, as the number of explanatory variables increases in a model, the adjusted R^2 becomes increasingly smaller than the unadjusted R^2. There seems to be a "penalty" involved in adding more explanatory variables to a regression model.
2. Although the unadjusted R^2 is always positive, the adjusted R^2 can on occasion turn out to be negative. For example, in a regression model involving $k = 3$ and $n = 30$, if an R^2 is found to be 0.06, R^2 can be negative (-0.0096).

At present, most computer regression packages compute both the adjusted and unadjusted R^2 values. This is a good practice, for the adjusted R^2 will enable us to compare two regressions that have the *same dependent variable* but a different number of explanatory variables.[15] Even when we are not comparing two regression models, it is a good practice to find the adjusted R^2 value because it explicitly takes into account the number of variables included in a model.

For our illustrative example, you should verify that the adjusted R^2 value is 0.8830, which, as expected, is smaller than the unadjusted R^2 value of 0.8906. The adjusted R^2 values for regressions (4.52) and (4.53) are 0.5169 and 0.1268, respectively, which are slightly lower than the corresponding unadjusted R^2 values.

4.11 WHEN TO ADD AN ADDITIONAL EXPLANATORY VARIABLE TO A MODEL

In practice, in order to explain a particular phenomenon, we are often faced with the problem of deciding among several competing explanatory variables. The common practice is to add variables as long as the adjusted R^2 increases (even though its numerical value may be smaller than the unadjusted R^2). But when does adjusted R^2 increase? *It can be shown that* $\overline{R}^2$ *will increase if the* $|t|$ *(absolute t) value of the coefficient of the added variable is larger than 1, where the t value is computed under the null hypothesis that the population value of the said coefficient is zero.*[16]

To see this all clearly, let us first regress auction price on a constant only, then on a constant and the age of the clock, and then on a constant, the age of the clock, and the number of bidders. The results are given in Table 4-4.

[15] As we will see in Chapter 5, if two regressions have different dependent variables, we cannot compare their R^2 values directly, adjusted or unadjusted.

[16] Whether or not a particular t value is significant, the adjusted R^2 will increase so long as the $|t|$ of the coefficient of the added variable is greater than 1.

TABLE 4-4 A COMPARISON OF FOUR MODELS OF ANTIQUE CLOCK AUCTION PRICES

Dependent variable	Intercept	Age	# of Bidders	R^2	$\overline{R}^2$	F	
Auction price	1328.094	—	—	0.00	0.00	0	(1)
	(19.0850)						
Auction price	−191.6662	10.4856	—	0.5325	0.5169	34.1723	(2)
	(−0.7248)	(5.8457)					
Auction price	807.9501	—	54.5724	0.1549	0.1268	5.5017	(3)
	(3.4962)		(2.3455)				
Auction price	−1336.049	12.7413	85.7640	0.8906	0.8830	118.0585	(4)
	(−7.6226)	(13.9653)	(9.7437)				

Note: Figures in the parentheses are the estimated t values under the null hypothesis that the corresponding population values are zero.

Some interesting facts stand out in this exercise:

1. When we regress auction price on the intercept only, the R^2, $\overline{R}^2$, *and* F values are all zero, as we would expect. But what does the intercept value represent here? It is nothing but the (sample) mean value of auction price. One way to check on this is to look at Eq. (2.16). If there is no X variable in this equation, the intercept is equal to the mean value of the dependent variable.
2. When we regress auction price on a constant and the age of the antique clock, we see that the t value of the age variable is not only greater than 1, but it is also statistically significant. Unsurprisingly, both R^2 *and* $\overline{R}^2$ values increase (although the latter is somewhat smaller than the former). But notice an interesting fact. If you square the t value of 5.8457, we get $(5.8457)^2 = 34.1722$, which is about the same as the F value of 34.1723 shown in Table 4-4. Is this surprising? No, because in Equation (C.15) in Appendix C we state that

$$t_k^2 = F_{1,k} \qquad \textbf{(4.55) = (C.15)}$$

   That is, the square of the t statistic with k d.f. is equal to the F statistic with 1 d.f. in the numerator and k d.f. in the denominator. In our example, $k = 30$ (32 observations − 2, the two coefficients estimated in model [2]). The numerator d.f. is 1, because we have only one explanatory variable in this model.
3. When we regress auction price on a constant and the number of bidders, we see that the t value of the latter is 2.3455. If you square this value, you will get $(2.3455)^2 = 5.5013$, which is about the same as the F value shown in Table 4-4, which again verifies Eq. (4.55). Since the t value is greater than 1, both R^2 *and* $\overline{R}^2$ values have increased. The computed t value is also statistically significant, suggesting that the number of bidders variable should be added to model (1). A similar conclusion holds for model (2).

4. How do we decide if it is worth adding both age and number of bidders together to model (1)? We have already answered this question with the help of the ANOVA technique and the attendant F test. In Table 4.2 we showed that one could reject the hypothesis that $B_2 = B_3 = 0$; that is, the two explanatory variables together have no impact on the auction bid price.[17]

4.12 RESTRICTED LEAST SQUARES

Let us take another look at the regressions given in Table 4-4. There we saw the consequences of omitting relevant variables from a regression model. Thus, in regression (1) shown in this table we regressed antique clock auction price on the intercept only, which gave an R^2 value of 0, which is not surprising. Then in regression (4) we regressed auction price on the age of the antique clock as well as on the number of bidders present at the auction, which gave an R^2 value of 0.8906. On the basis of the F test we concluded that there was a specification error and that both the explanatory variables should be added to the model.

Let us call regression (1) the *restricted model* because it implicitly assumes that the coefficients of the age of the clock and the number of bidders are zero; that is, these variables do not belong in the model (i.e., $B_2 = B_3 = 0$). Let us call regression (4) the *unrestricted model* because it includes all the relevant variables. Since (1) is a restricted model, when we estimate it by OLS, we call it **restricted least squares (RLS).** Since (4) is an unrestricted model, when we estimate it by OLS, we call it **unrestricted least squares (URLS).** All the models we have estimated thus far have been essentially URLS, for we have assumed that the model being estimated has been correctly specified and that we have included all the relevant variables in the model. In Chapter 7 we will see the consequences of violating this assumption.

The question now is: How do we decide between RLS and URLS? That is, how do we find out if the restrictions imposed by a model, such as (1) in the present instance, are valid? This question can be answered by the F test. For this purpose, let R_r^2 denote the R^2 value obtained from the restricted model and R_{ur}^2 denote the R^2 value obtained from the unrestricted model. Now assuming that the error term u_i is normally distributed, it can be shown that

$$F = \frac{\left(R_{ur}^2 - R_r^2\right)/m}{\left(1 - R_{ur}^2\right)/(n-k)} \sim F_{m,n-k} \tag{4.56}$$

follows the F distribution with m and $(n-k)$ d.f. in the numerator and denominator, respectively, where $R_r^2 = R^2$ obtained from the restricted regression,

[17]Suppose you have a model with four explanatory variables. Initially you only include two of these variables but then you want to find out if it is worth adding two more explanatory variables. This can be handled by an extension of the F test. For details, see Gujarati and Porter, *Basic Econometrics*, 5th ed., McGraw-Hill, New York, 2009, pp. 243–246.

$R^2_{ur} = R^2$ obtained from the unrestricted regression, m = number of restrictions imposed by the restricted regression (two in our example), n = number of observations in the sample, and k = number of parameters estimated in the unrestricted regression (including the intercept). The null hypothesis tested here is that the restrictions imposed by the restricted model are valid. If the F value estimated from Equation (4.56) exceeds the critical F value at the chosen level of significance, we reject the restricted regression. That is, in this situation, the restrictions imposed by the (restricted) model are not valid.

Returning to our antique clock auction price example, putting the appropriate values in Eq. (4.56) from Table 4-4, we obtain:

$$F = \frac{(0.890 - 0)/2}{(1 - 0.890)/(32 - 3)} = \frac{0.445}{0.00379} = 117.414 \qquad \textbf{(4.57)}$$

The probability of such an F value is extremely small. Therefore, we reject the restricted regression. More positively, age of the antique clock as well as the number of bidders at auction both have a statistically significant impact on the auction price.

The formula (4.56) is of general application. The only precaution to be taken in its application is that in comparing the restricted and unrestricted regressions, the dependent variables must be in the same form. If they are not, we have to make them comparable using the method discussed in Chapter 5 (see Problem 5.16) or use an alternative that is discussed in Exercise 4.20.

4.13 ILLUSTRATIVE EXAMPLES

To conclude this chapter, we consider several examples involving multiple regressions. Our objective here is to show you how multiple regression models are used in a variety of applications.

Example 4.1. Does Tax Policy Affect Corporate Capital Structure?

To find out the extent to which tax policy has been responsible for the recent trend in U.S. manufacturing toward increasing use of debt capital in lieu of equity capital—that is, toward an increasing debt/equity ratio (called leverage in the financial literature)—Pozdena estimated the following regression model:[18]

$$Y_t = B_1 + B_2 X_{2t} + B_3 X_{3t} + B_4 X_{4t} + B_5 B X_{5t} + B_6 X_{6t} + u_t \qquad \textbf{(4.58)}$$

where Y = the leverage (= debt/equity) in percent
X_2 = the corporate tax rate
X_3 = the personal tax rate
X_4 = the capital gains tax rate
X_5 = nondebt-related tax shields
X_6 = the inflation rate

[18]Randall Johnston Pozdena, "Tax Policy and Corporate Capital Structure," *Economic Review,* Federal Reserve Bank of San Francisco, Fall 1987, pp. 37–51.

TABLE 4-5 LEVERAGE IN MANUFACTURING CORPORATIONS, 1935–1982

Explanatory variable	Coefficient (t value in parentheses)
Corporate tax rate	2.4
	(10.5)
Personal tax rate	−1.2
	(−4.8)
Capital gains tax rate	0.3
	(1.3)
Non-debt-related tax shield	−2.4
	(−4.8)
Inflation rate	1.4
	(3.0)
$n = 48$ (number of observations)	
$R^2 = 0.87$	
$\bar{R}^2 = 0.85$	

Notes: 1. The author does not present the estimated intercept.
2. The adjusted R^2 is calculated using Eq. (4.54).
3. The standard errors of the various coefficients can be obtained by dividing the coefficient value by its t value (e.g., 2.4/10.5 = 0.2286 is the se of the corporate tax rate coefficient).

Source: Randall Johnston Pozdena, "Tax Policy and Corporate Capital Structure," *Economic Review,* Federal Reserve Bank of San Francisco, Fall 1987, Table 1, p. 45 (adapted).

Economic theory suggests that coefficients B_2, B_4, and B_6 will be positive and coefficients B_3 and B_5 will be negative.[19] Based on the data for U.S. manufacturing corporations for the years 1935 to 1982, Pozdena obtained the OLS results that are presented in tabular form (Table 4-5) rather than in the usual format (e.g., Eq. [4.37]). (Results are sometimes presented in this form for ease of reading.)

Discussion of Regression Results

The first fact to note about the preceding regression results is that all the coefficients have signs according to prior expectations. For instance, the corporate tax rate has a positive effect on leverage. Holding other things the same, as the corporate tax rate goes up by 1 percentage point, on the average, the leverage ratio (i.e., the debt/equity ratio) goes up by 2.4 percentage points. Likewise, if the inflation rate goes up by 1 percentage point, on the average, leverage goes up by 1.4 percentage points, other things remaining the same. (*Question:* Why would you expect a positive relation between leverage and inflation?) Other partial regression coefficients should be interpreted similarly.

Since the t values are presented underneath each partial regression coefficient under the null hypothesis that each population partial regression coefficient is

[19]See Pozdena's article (footnote 18) for the theoretical discussion of expected signs of the various coefficients. In the United States the interest paid on debt capital is tax deductible, whereas the income paid as dividends is not. This is one reason that corporations may prefer debt to equity capital.

individually equal to zero, we can easily test whether such a null hypothesis stands up against the (two-sided) alternative hypothesis that each true population coefficient is different from zero. Hence, we use the two-tail t test. The d.f. in this example are 42, which are obtained by subtracting from n (= 48) the number of parameters estimated, which are 6 in the present instance. (*Note:* The intercept value is not presented in Table 4-5, although it was estimated.) If we choose $\alpha = 0.05$ or 5%, the two-tail *critical t value* is about 2.021 for 40 d.f. (*Note:* This is good enough for present purposes since the t table does not give the precise t value for 42 d.f.) If α is fixed at 0.01 or a 1% level, the critical t value for 40 d.f. is 2.704 (two-tail). Looking at the t values presented in Table 4-5, we see that each partial regression coefficient, except that of the capital gains tax variable, is statistically significantly different from zero at the 1% level of significance. The coefficient of the capital gains tax variable is not significant at either the 1% or 5% level. Therefore, except for this variable, we can reject the *individual* null hypothesis that each partial regression coefficient is zero. In other words, all but one of the explanatory variables *individually* has an impact on the debt/equity ratio. In passing, note that *if an estimated coefficient is statistically significant at the 1% level, it is also significant at the 5% level, but the converse is not true.*

What about the overall significance of the estimated regression line? That is, do we reject the null hypothesis that all partial slopes are *simultaneously* equal to zero or, equivalently, is $R^2 = 0$? This hypothesis can be easily tested by using Eq. (4.50), which in the present case gives

$$\begin{aligned} F &= \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \\ &= \frac{0.87/5}{0.13/42} \\ &= 56.22 \end{aligned} \tag{4.59}$$

This F value has an F distribution with 5 and 42 d.f. If α is set at 0.05, the F table (Appendix E, Table E-3) shows that for 5 and 40 d.f. (the table has no exact value of 42 d.f. in the denominator), the *critical F value* is 2.45. The corresponding value at $\alpha = 0.01$ is 3.51. The computed F of ≈ 56 far exceeds either of these critical F values. Therefore, we reject the null hypothesis that all partial slopes are simultaneously equal to zero or, alternatively, $R^2 = 0$. *Collectively,* all five explanatory variables influence the dependent variable. *Individually,* however, as we have seen, only four variables have an impact on the dependent variable, the debt/equity ratio. Example 4.1 again underscores the point made earlier that the (individual) t test and the (joint) F test are quite different.[20]

[20]In the two-variable linear regression model, as noted before, $t_k^2 = F_{1,k}$; that is, the square of a t value with k d.f. is equal to an F value with 1 d.f. in the numerator and k d.f. in the denominator.

Example 4.2. The Demand for Imports in Jamaica

To explain the demand for imports in Jamaica, J. Gafar[21] obtained the following regression based on annual data for 19 years:

$$\begin{aligned} \hat{Y}_t &= -58.9 + 0.20X_{2t} - 0.10X_{3t} \\ \text{se} &= \quad (0.0092) \quad (0.084) \qquad R^2 = 0.96 \\ t &= \quad (21.74) \quad (-1.1904) \qquad \bar{R}^2 = 0.955 \end{aligned} \tag{4.60}$$

where Y = quantity of imports
X_2 = personal consumption expenditure
X_3 = import price/domestic price

Economic theory would suggest a positive relationship between Y and X_2 and a negative relationship between Y and X_3, which turns out to be the case. Individually, the coefficient of X_2 is statistically significant but that of X_3 is not at, say, the 5% level. But since the absolute t value of X_3 is greater than 1, $\bar{R}^2$ for this example will drop if X_3 is dropped from the model. (Why?) Together, X_2 and X_3 explain about 96 percent of the variation in the quantity of imports into Jamaica.

Example 4.3. The Demand for Alcoholic Beverages in the United Kingdom

To explain the demand for alcoholic beverages in the United Kingdom, T. McGuinness[22] estimated the following regression based on annual data for 20 years:

$$\begin{aligned} \hat{Y}_t &= -\ 0.014 - 0.354X_{2t} + 0.0018X_{3t} + 0.657X_{4t} + 0.0059X_{5t} \\ \text{se} &= (0.012) \quad (0.2688) \quad (0.0005) \quad (0.266) \quad (0.0034) \\ t &= (-1.16) \quad (1.32) \quad (3.39) \quad (2.47) \quad (1.73) \\ R^2 &= 0.689 \end{aligned} \tag{4.61}$$

where Y = the annual change in pure alcohol consumption per adult
X_2 = the annual change in the real price index of alcoholic drinks
X_3 = the annual change in the real disposable income per person
$X_4 = \dfrac{\text{the annual change in the number of licensed premises}}{\text{the adult population}}$
X_5 = the annual change in real advertising expenditure on alcoholic drinks per adult

Theory would suggest that all but the variable X_2 will be positively related to Y. This is borne out by the results, although not all coefficients are

[21] J. Gafar, "Devaluation and the Balance of Payments Adjustment in a Developing Economy: An Analysis Relating to Jamaica," *Applied Economics*, vol. 13, 1981, pp. 151–165. Notations were adapted. Adjusted R^2 computed.

[22] T. McGuinness, "An Econometric Analysis of Total Demand for Alcoholic Beverages in the United Kingdom," *Journal of Industrial Economics*, vol. 29, 1980, pp. 85–109. Notations were adapted.

individually statistically significant. For 15 d.f. (Why?), the 5% critical t value is 1.753 (one-tail) and 2.131 (two-tail). Consider the coefficient of X_5, the change in advertising expenditure. Since the advertising expenditure and the demand for alcoholic beverages are expected to be positive (otherwise, it is bad news for the advertising industry), we can entertain the hypothesis that $H_0: B_5 = 0$ vs. $H_1: B_5 > 0$ and therefore use the one-tail t test. The computed t value of 1.73 is very close to being significant at the 5% level.

It is left as an exercise for you to compute the F value for this example to test the hypothesis that all partial slope coefficients are simultaneously equal to zero.

Example 4.4. Civilian Labor Force Participation Rate, Unemployment Rate, and Average Hourly Earnings Revisited

In Chapter 1 we presented regression (1.5) without discussing the statistical significance of the results. Now we have the necessary tools to do that. The complete regression results are as follows:

$$\begin{aligned} \widehat{CLFPR}_t &= 81.2267 - 0.6384 CUNR_t - 1.4449 AHE82_t \\ \text{se} &= (3.4040) \quad (0.0715) \quad (0.4148) \\ t &= (23.88) \quad (-8.94) \quad (-3.50) \\ p \text{ value} &= (0.000)^* \quad (0.000)^* \quad (0.002) \\ R^2 &= 0.767; \ \overline{R}^2 = 0.748; \quad F = 41.09 \end{aligned} \tag{4.62}$$

As these results show, each of the estimated regression coefficients is individually statistically highly significant, because the p values are so small. That is, each coefficient is significantly different from zero. Collectively, both *CUNR* and *AHE82* are also highly statistically significant, because the p value of the computed F value (for 2 and 25 d.f.) of 41.09 is extremely low.

As expected, the civilian unemployment rate has a negative relationship to the civilian labor force participation rate, suggesting that perhaps the discouraged-worker effect dominates the added-worker hypothesis. The theoretical reasoning behind this has already been explained in Chapter 1. The negative value of *AHE82* suggests that perhaps the income effect dominates the substitution effect.

Example 4.5. Expenditure on Education in 38 Countries:[23]

Based on data taken from a sample of 38 countries (see Table 4-6, found on the textbook's Web site), we obtained the following regression:

$$\text{Educ}_i = 414.4583 + 0.0523 GDP_i - 50.0476\ Pop$$

[23]The data used in this exercise are from Gary Koop, *Introduction to Econometrics*, John Wiley & Sons, England, 2008 and can be found on the following Web site: **www.wileyeurope.com/college/koop**.

*Denotes extremely small value.

$$\begin{aligned} \text{se} &= (266.4583) \quad (0.0018) \quad (9.9581) \\ t &= (1.5538) \quad (28.2742) \quad (-5.0257) \\ p \text{ value} &= (0.1292) \quad (0.0000) \quad (0.0000) \end{aligned}$$

$$R^2 = 0.9616;\ \overline{R}^2 = 0.9594;\ F = 439.22;\ p \text{ value of } F = 0.000$$

where Educ = expenditure on education (millions of U.S. dollars), GDP = gross domestic product (millions of U.S. dollars), and Pop = population (millions of people). As you can see from the data, the sample includes a variety of countries in different stages of economic development.

It can be readily assessed that the GDP and Pop variables are individually highly significant, although the sign of the population variable may be puzzling. Since the estimated F is so highly significant, collectively the two variables have a significant impact on expenditure on education. As noted, the variables are also individually significant.

The R^2 and adjusted $\overline{R}^2$ square values are quite high, which is unusual in a cross-section sample of diverse countries.

We will explore these data further in later chapters.

4.14 SUMMARY

In this chapter we considered the simplest of the multiple regression models, namely, the three-variable linear regression model—one dependent variable and two explanatory variables. Although in many ways a straightforward extension of the two-variable linear regression model, the three-variable model introduced several new concepts, such as partial regression coefficients, adjusted and unadjusted multiple coefficient of determination, and multicollinearity.

Insofar as estimation of the parameters of the multiple regression coefficients is concerned, we still worked within the framework of the classical linear regression model and used the method of ordinary least squares (OLS). The OLS estimators of multiple regression, like the two-variable model, possess several desirable statistical properties summed up in the Gauss-Markov property of best linear unbiased estimators (BLUE).

With the assumption that the disturbance term follows the normal distribution with zero mean and constant variance σ^2, we saw that, as in the two-variable case, each estimated coefficient in the multiple regression follows the normal distribution with a mean equal to the true population value and the variances given by the formulas developed in the text. Unfortunately, in practice, σ^2 is not known and has to be estimated. The OLS estimator of this unknown variance is $\hat{\sigma}^2$. But if we replace σ^2 by $\hat{\sigma}^2$, then, as in the two-variable case, each estimated coefficient of the multiple regression follows the t distribution, not the normal distribution.

The knowledge that each multiple regression coefficient follows the t distribution with d.f. equal to $(n - k)$, where k is the number of parameters estimated (including the intercept), means we can use the t distribution to test

statistical hypotheses about each multiple regression coefficient individually. This can be done on the basis of either the t test of significance or the confidence interval based on the t distribution. In this respect, the multiple regression model does not differ much from the two-variable model, except that proper allowance must be made for the d.f., which now depend on the number of parameters estimated.

However, when testing the hypothesis that all partial slope coefficients are simultaneously equal to zero, the individual t testing referred to earlier is of no help. Here we should use the analysis of variance (ANOVA) technique and the attendant F test. Incidentally, testing that all partial slope coefficients are simultaneously equal to zero is the same as testing that the multiple coefficient of determination R^2 is equal to zero. Therefore, the F test can also be used to test this latter but equivalent hypothesis.

We also discussed the question of when to add a variable or a group of variables to a model, using either the t test or the F test. In this context we also discussed the method of restricted least squares.

All the concepts introduced in this chapter have been illustrated by numerical examples and by concrete economic applications.

KEY TERMS AND CONCEPTS

The key terms and concepts introduced in this chapter are

Multiple regression model
  Partial regression coefficients; partial slope coefficients
Multicollinearity
Collinearity; exact linear relationship
  a) high or near perfect collinearity
Multiple coefficient of determination, R^2
Coefficient of multiple correlation, R
Individual hypothesis testing
Joint hypothesis testing or test of overall significance of estimated multiple regression
  a) analysis of variance (ANOVA)
  b) F test
Model specification bias (specification error)
Adjusted R^2 ($\overline{R}^2$)
Restricted least squares (RLS)
Unrestricted least squares (URLS)
Relationship between t and F tests

QUESTIONS

4.1. Explain carefully the meaning of
  a. Partial regression coefficient
  b. Coefficient of multiple determination, R^2
  c. Perfect collinearity
  d. Perfect multicollinearity
  e. Individual hypothesis testing
  f. Joint hypothesis testing
  g. Adjusted R^2

4.2. Explain step by step the procedure involved in
- **a.** Testing the statistical significance of a single multiple regression coefficient.
- **b.** Testing the statistical significance of all partial slope coefficients.

4.3. State with brief reasons whether the following statements are true (T), false (F), or uncertain (U).
- **a.** The adjusted and unadjusted R^2s are identical only when the unadjusted R^2 is equal to 1.
- **b.** The way to determine whether a group of explanatory variables exerts significant influence on the dependent variable is to see if any of the explanatory variables has a significant t statistic; if not, they are statistically insignificant as a group.
- **c.** When $R^2 = 1$, $F = 0$, and when $R^2 = 0$, $F =$ infinite.
- **d.** When the d.f. exceed 120, the 5% critical t value (two-tail) and the 5% critical Z (standard normal) value are identical, namely, 1.96.
- ***e.** In the model $Y_i = B_1 + B_2X_{2i} + B_3X_{3i} + u_i$, if X_2 and X_3 are negatively correlated in the sample and $B_3 > 0$, omitting X_3 from the model will bias b_{12} downward [i.e., $E(b_{12}) < B_2$] where b_{12} is the slope coefficient in the regression of Y on X_2 alone.
- **f.** When we say that an estimated regression coefficient is statistically significant, we mean that it is statistically different from 1.
- **g.** To compute a critical t value, we need to know only the d.f.
- **h.** By the overall significance of a multiple regression we mean the statistical significance of any single variable included in the model.
- **i.** Insofar as estimation and hypothesis testing are concerned, there is no difference between simple regression and multiple regression.
- **j.** The d.f. of the total sum of squares (TSS) are always $(n - 1)$ regardless of the number of explanatory variables included in the model.

4.4. What is the value of $\hat{\sigma}^2$ in each of the following cases?
- **a.** $\sum e_i^2 = 880$, $n = 25$, $k = 4$ (including intercept)
- **b.** $\sum e_i^2 = 1220$, n = 14, k = 3 (excluding intercept)

4.5. Find the critical t value(s) in the following situations:

Degrees of freedom (d.f.)	Level of significance (%)	H_0
12	5	Two-tail
20	1	Right-tail
30	5	Left-tail
200	5	Two-tail

4.6. Find the critical F values for the following combinations:

Numerator d.f.	Denominator d.f.	Level of significance (%)
5	5	5
4	19	1
20	200	5

* Optional.

PROBLEMS

4.7. You are given the following data:

Y	X_2	X_3
1	1	2
3	2	1
8	3	−3

Based on these data, estimate the following regressions *(Note:* Do not worry about estimating the standard errors):

a. $Y_i = A_1 + A_2X_{2i} + u_i$
b. $Y_i = C_1 + C_3X_{3i} + u_i$
c. $Y_i = B_1 + B_2X_{2i} + B_3X_{3i} + u_i$
d. Is $A_2 = B_2$? Why or why not?
e. Is $C_3 = B_3$? Why or why not?

What general conclusion can you draw from this exercise?

4.8. You are given the following data based on 15 observations:

$$\overline{Y} = 367.693; \quad \overline{X}_2 = 402.760; \quad \overline{X}_3 = 8.0; \quad \sum y_i^2 = 66{,}042.269$$
$$\sum x_{2i}^2 = 84{,}855.096; \quad \sum x_{3i}^2 = 280.0; \quad \sum y_i x_{2i} = 74{,}778.346$$
$$\sum y_i x_{3i} = 4{,}250.9; \quad \sum x_{2i} x_{3i} = 4{,}796.0$$

where lowercase letters, as usual, denote deviations from sample mean values.

a. Estimate the three multiple regression coefficients.
b. Estimate their standard errors.
c. Obtain R^2 and $\overline{R}^2$.
d. Estimate 95% confidence intervals for B_2 and B_3.
e. Test the statistical significance of each estimated regression coefficient using $\alpha = 5\%$ (two-tail).
f. Test at $\alpha = 5\%$ that all partial slope coefficients are equal to zero. Show the ANOVA table.

4.9. A three-variable regression gave the following results:

Source of variation	Sum of squares (SS)	d.f.	Mean sum of squares (MSS)
Due to regression (ESS)	65,965	—	—
Due to residual (RSS)	—	—	—
Total (TSS)	66,042	14	

a. What is the sample size?
b. What is the value of the RSS?
c. What are the d.f. of the ESS and RSS?
d. What is R^2? And $\overline{R}^2$?
e. Test the hypothesis that X_2 and X_3 have zero influence on Y. Which test do you use and why?
f. From the preceding information, can you determine the individual contribution of X_2 and X_3 toward Y?

4.10. Recast the ANOVA table given in problem 4.9 in terms of R^2.

4.11. To explain what determines the price of air conditioners, B. T. Ratchford[24] obtained the following regression results based on a sample of 19 air conditioners:

$$\hat{Y}_i = -68.236 + 0.023X_{2i} + 19.729X_{3i} + 7.653X_{4i} \quad R^2 = 0.84$$

$$\text{se} = \qquad (0.005) \qquad (8.992) \qquad (3.082)$$

where Y = the price, in dollars
X_2 = the BTU rating of air conditioner
X_3 = the energy efficiency ratio
X_4 = the number of settings
se = standard errors

a. Interpret the regression results.
b. Do the results make economic sense?
c. At $\alpha = 5\%$, test the hypothesis that the BTU rating has no effect on the price of an air conditioner versus that it has a positive effect.
d. Would you accept the null hypothesis that the three explanatory variables explain a substantial variation in the prices of air conditioners? Show clearly all your calculations.

4.12. Based on the U.S. data for 1965-IQ to 1983-IVQ ($n = 76$), James Doti and Esmael Adibi[25] obtained the following regression to explain personal consumption expenditure (PCE) in the United States.

$$\hat{Y}_t = -10.96 + 0.93X_{2t} - 2.09X_{3t}$$

$$t = (-3.33) \quad (249.06) \quad (-3.09) \qquad R^2 = 0.9996$$

$$F = 83{,}753.7$$

where Y = the PCE ($, in billions)
X_2 = the disposable (i.e., after-tax) income ($, in billions)
X_3 = the prime rate (%) charged by banks

a. What is the marginal propensity to consume (MPC)—the amount of additional consumption expenditure from an additional dollar's personal disposable income?
b. Is the MPC statistically different from 1? Show the appropriate testing procedure.
c. What is the rationale for the inclusion of the prime rate variable in the model? A priori, would you expect a negative sign for this variable?
d. Is b_3 significantly different from zero?
e. Test the hypothesis that $R^2 = 0$.
f. Compute the standard error of each coefficient.

[24]B. T. Ratchford, "The Value of Information for Selected Appliances," *Journal of Marketing Research*, vol. 17, 1980, pp. 14–25. Notations were adapted.
[25]James Doti and Esmael Adibi, *Econometric Analysis: An Applications Approach*, Prentice-Hall, Englewood Cliffs, N.J., 1988, p. 188. Notations were adapted.

4.13. In the illustrative Example 4.2 given in the text, test the hypothesis that X_2 and X_3 together have no influence on Y. Which test will you use? What are the assumptions underlying that test?

4.14. Table 4-7 (found on the textbook's Web site) gives data on child mortality (CM), female literacy rate (FLR), per capita GNP (PGNP), and total fertility rate (TFR) for a group of 64 countries.

- **a.** A priori, what is the expected relationship between CM and each of the other variables?
- **b.** Regress CM on FLR and obtain the usual regression results.
- **c.** Regress CM on FLR and PGNP and obtain the usual results.
- **d.** Regress CM on FLR, PGNP, and TFR and obtain the usual results. Also show the ANOVA table.
- **e.** Given the various regression results, which model would you choose and why?
- **f.** If the regression model in (*d*) is the correct model, but you estimate (*a*) or (*b*) or (*c*), what are the consequences?
- **g.** Suppose you have regressed CM on FLR as in (*b*). How would you decide if it is worth adding the variables PGNP and TFR to the model? Which test would you use? Show the necessary calculations.

4.15. Use formula (4.54) to answer the following question:

Value of R^2	n	k	$\overline{R}^2$
0.83	50	6	—
0.55	18	9	—
0.33	16	12	—
0.12	1,200	32	—

What conclusion do you draw about the relationship between R^2 and $\overline{R}^2$?

4.16. For Example 4.3, compute the F value. If that F value is significant, what does that mean?

4.17. For Example 4.2, set up the ANOVA table and test that $R^2 = 0$. Use $\alpha = 1\%$.

4.18. Refer to the data given in Table 2-12 (found on the textbook's Web site) to answer the following questions:

- **a.** Develop a multiple regression model to explain the average starting pay of MBA graduates, obtaining the usual regression output.
- **b.** If you include both GPA and GMAT scores in the model, a priori, what problem(s) may you encounter and why?
- **c.** If the coefficient of the tuition variable is positive and statistically significant, does that mean it pays to go to the most expensive business school? What might the tuition variable be a proxy for?
- **d.** Suppose you regress GMAT score on GPA and find a statistically significant positive relationship between the two. What can you say about the problem of multicollinearity?
- **e.** Set up the ANOVA table for the multiple regression in part (*a*) and test the hypothesis that all partial slope coefficients are zero.
- **f.** Do the ANOVA exercise in part (*e*), using the R^2 value.

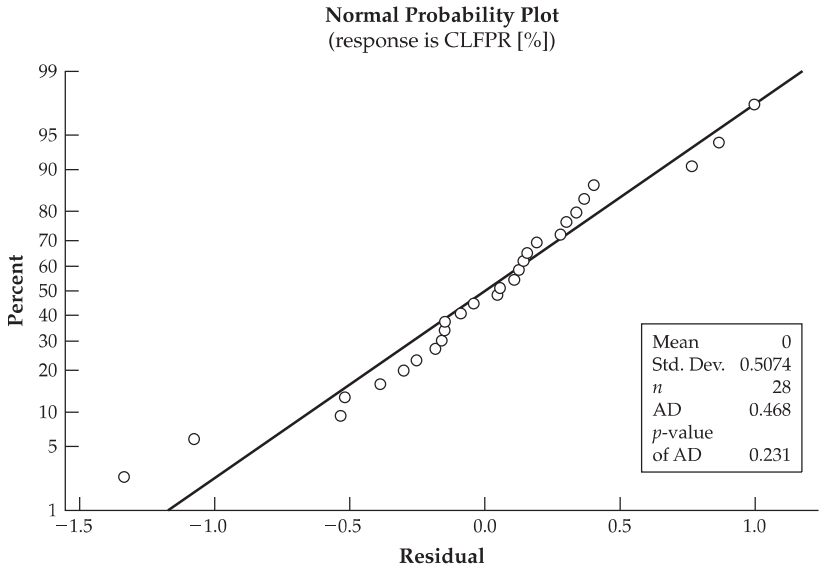

FIGURE 4-1 Normal probability plot for Example 4.4
AD = Anderson-Darling statistic

4.19. Figure 4-1 gives you the normal probability plot for Example 4.4.
- **a.** From this figure, can you tell if the error term in Eq. (4.62) follows the normal distribution? Why or why not?
- **b.** Is the observed Anderson-Darling A^2 value of 0.468 statistically significant? If it is, what does that mean? If it is not, what conclusion do you draw?
- **c.** From the given data, can you identify the mean and variance of the error term?

4.20. *Restricted least squares (RLS).* If the dependent variables in the restricted and unrestricted regressions are not the same, you can use the following variant of the F test given in Eq. (4.56)

$$F = \frac{(RSS_r - RSS_{ur})/m}{RSS_{ur}/(n-k)} \sim F_{m,n-k}$$

where RSS_r = residual sum of squares from the restricted regression, RSS_{ur} = residual sum of squares from the unrestricted regression, m = number of restrictions, and $(n - k)$ = d.f. in the unrestricted regression.

Just to familiarize yourself with this formula, rework the model given in Table 4-4.

4.21. Refer to Example 4.5.
- **a.** Use the method of restricted least squares to find out if it is worth adding the Pop (population) variable to the model.
- **b.** Divide both Educ and GDP by Pop to obtain per capita Educ and per capita GDP. Now regress per capita Educ on per capita GDP and compare your

results with those given in Example 4.5. What conclusion can you draw from this exercise?

4.22. Table 4-8 (found on the textbook's Web site) contains variables from the Los Angeles 2008 Zagat Restaurant Guide. The variables are score values out of 30, with 30 being the best. For each restaurant listed, the table provides data for four categories: food, décor, service, and average price for a single meal at the establishment.

a. Create a least squares regression model to predict Price based on the other three variables (Food, Décor, and Service). Are all the independent variables statistically significant?

b. Does the normal probability plot indicate any problems?

c. Create a scattergram of the residual values from the model versus the fitted values of the Price estimates. Does the plot indicate the residual values have constant variance? Retain this plot for use in future chapters.

APPENDIX 4A.1: Derivations of OLS Estimators Given in Equations (4.20) to (4.22)

Start with Eq. (4.16). Differentiate this equation partially with respect to b_1, b_2, and b_3, and set the resulting equations to zero to obtain:

$$\frac{\partial \sum e_i^2}{\partial \sum b_1} = 2\sum(Y_i - b_1 - b_2X_{2i} - b_3X_{3i})(-1) = 0$$

$$\frac{\partial \sum e_i^2}{\partial b_2} = 2\sum(Y_i - b_1 - b_2X_2 - b_3X_{3i})(-X_{2i}) = 0$$

$$\frac{\partial \sum e_i^2}{\partial b_3} = 2\sum(Y_i - b_1 - b_2X_{2i} - b_3X_3)(-X_{3i}) = 0$$

Simplifying these equations gives Eq. (4.17), (4.18), and (4.19). Using small letters to denote deviations from the mean values (e.g., $x_{2i} = X_{2i} - \overline{X}_2$), we can solve the preceding equations to obtain the formulas given in Eqs. (4.20), (4.21), and (4.22).

APPENDIX 4A.2: Derivation of Equation (4.31)

Note that the three-variable sample regression model

$$Y_i = b_1 + b_2X_{2i} + b_3X_{3i} + e_i \qquad \textbf{(4A.2.1)}$$

can be expressed in the deviation form (i.e., each variable expressed as a deviation from the mean value and noting that $\bar{e} = 0$) as

$$y_i = b_2x_{2i} + b_3x_{3i} + e_i \qquad \textbf{(4A.2.2)}$$

Therefore,

$$e_i = y_i - b_2 x_{2i} - b_3 x_{3i} \tag{4A.2.3}$$

Which we can write as

$$\begin{aligned}
\sum e_i^2 &= \sum (e_i e_i) \\
&= \sum e_i (y_i - b_2 x_{2i} - b_3 x_{3i}) \\
&= \sum e_i y_i - b_2 \sum e_i x_{2i} - b_3 \sum e_i x_{3i} \\
&= \sum e_i y_i \qquad \text{since the last two terms are zero (why?)} \\
&= \sum (y_i - b_2 x_{2i} - b_3 x_{3i})(y_i) \\
&= \sum y_i^2 - b_2 \sum y_i x_{2i} - b_3 \sum y_i x_{3i} \\
&= \sum y_i^2 - (b_2 \sum y_i x_{2i} + b_3 \sum y_i x_{3i}) \\
&= \text{TSS} - \text{ESS}
\end{aligned}$$

APPENDIX 4A.3: Derivation of Equation (4.50)

Recall that (see footnote 9)

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{4A.3.1}$$

Now $\overline{R}^2$ is defined as

$$\begin{aligned}
\overline{R}^2 &= 1 - \frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)} \\
&= 1 - \frac{\text{RSS}\,(n-1)}{\text{TSS}\,(n-k)}
\end{aligned} \tag{4A.3.2}$$

Note how the degrees of freedom are taken into account.

Now substituting Equation (4A.3.1) into Equation (4A.3.2), and after algebraic manipulations, we obtain

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

Notice that if we do not take into account the d.f. associated with RSS ($= n - k$) and TSS ($= n - 1$), then, obviously $\overline{R}^2 = R^2$.

APPENDIX 4A.4: EViews Output of the Clock Auction Price Example

Method: Least Squares

Sample: 1 32
Included observations: 32

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	−1336.049	175.2725	−7.622698	0.0000
AGE	12.74138	0.912356	13.96537	0.0000
NOBID	85.76407	8.801995	9.743708	0.0000

R-squared	0.890614	Mean dependent var	1328.094
Adjusted R-squared	0.883070	S.D. dependent var	393.6495
S.E. of regression	134.6083	Akaike info criterion	12.73167
Sum squared resid	525462.2	Schwarz criterion	12.86909
Log likelihood	−200.7068	*F*-statistic	118.0585
Durbin-Watson stat	1.864656	Prob (*F*-statistic)	0.000000

Actual Y	Fitted $(\hat{Y})$	Residual e_i	Residual Plot
1235.00	1397.04	−162.039	
1080.00	1158.38	−78.3786	
845.000	882.455	−37.4549	
1552.00	1347.03	204.965	
1047.00	1166.19	−119.191	
1979.00	1926.29	52.7127	
1822.00	1680.78	141.225	
1253.00	1203.45	49.5460	
1297.00	1181.40	115.603	
946.000	875.604	70.3963	
1713.00	1695.98	17.0187	
1024.00	1098.10	−74.0973	
2131.00	2030.68	100.317	
1550.00	1669.00	−118.995	
1884.00	1671.46	212.540	
2041.00	1866.01	174.994	
854.000	1000.55	−146.553	
1483.00	1461.71	21.2927	
1055.00	1240.72	−185.717	
1545.00	1579.81	−34.8054	
729.000	554.605	174.395	
1792.00	1716.53	75.4650	
1175.00	1364.71	−189.705	
1593.00	1732.70	−139.702	
1147.00	1095.63	51.3672	
1092.00	1127.97	−35.9668	
1152.00	1269.63	−117.625	
1336.00	1127.01	208.994	
785.000	678.593	106.407	
744.000	729.558	14.4417	
1356.00	1564.60	−208.599	
1262.00	1404.85	−142.852