CHAPTER 3

THE TWO-VARIABLE MODEL: HYPOTHESIS TESTING

In Chapter 2 we showed how the method of least squares works. By applying that method to our math S.A.T. sample data given in Table 2-2, we obtained the following math S.A.T. score function:

$$\hat{Y}_i = 432.4138 + 0.0013X_i \qquad \textbf{(2.20)}$$

where Y represents math S.A.T. score and X represents annual family income, measured in dollars.

This example illustrated the estimation stage of statistical inference. We now turn our attention to its other stage, namely, hypothesis testing. The important question that we raise is: How "good" is the estimated regression line given in Equation (2.20)? That is, how can we tell that it really is a good estimator of the true population regression function (PRF)? How can we be sure just on the basis of a single sample given in Table 2-2 that the estimated regression function (i.e., the sample regression function [SRF]) is in fact a good approximation of the true PRF?

We cannot answer this question definitely unless we are a little more specific about our PRF, Eq. (2.2). As Eq. (2.2) shows, Y_i depends on both X_i and u_i. Now we have assumed that the X_i values are known or given—recall from Chapter 2 that our analysis is a conditional regression analysis, conditional upon the given X's. In short, we treat the X values as *nonstochastic.* The (nonobservable) error term u is of course random, or stochastic. (Why?) Since a stochastic term (u) is added to a nonstochastic term (X) to generate Y, Y becomes stochastic, too. This means that unless we are willing to assume how the stochastic u terms are generated, we will not be able to tell how good an SRF is as an estimate of the true PRF.

In deriving the ordinary least squares (OLS) estimators so far, we did not say how the u_i were generated, for the derivation of OLS estimators did not depend on any (probabilistic) assumption about the error term. But in testing statistical hypotheses based on the SRF, we cannot make further progress, as we will show shortly, unless we make some specific assumptions about how u_i are generated. This is precisely what the so-called **classical linear regression model (CLRM)** does, which we will now discuss. Again, to explain the fundamental ideas, we consider the two-variable regression model introduced in Chapter 2. In Chapter 4 we extend the ideas developed here to the multiple regression models.

3.1 THE CLASSICAL LINEAR REGRESSION MODEL

The CLRM makes the following assumptions:

A3.1.

The regression model is *linear in the parameters*; it may or may not be linear in the variables. That is, the regression model is of the following type.

$$Y_i = B_1 + B_2 X_i + u_i \qquad \textbf{(2.2)}$$

As will be discussed in Chapter 4, this model can be extended to include more explanatory variables.

A3.2.

The explanatory variable(s) X is uncorrelated with the disturbance term u. However, if the X variable(s) is *nonstochastic* (i.e., its value is a fixed number), this assumption is automatically fulfilled. Even if the X value(s) is stochastic, with a large enough sample size this assumption can be related without severely affecting the analysis.[1]

This assumption is not a new assumption because in Chapter 2 we stated that our regression analysis is a *conditional regression analysis*, conditional upon the given X values. In essence, we are assuming that the X's are nonstochastic. Assumption (3.1) is made to deal with simultaneous equation regression models, which we will discuss in Chapter 11.

A3.3.

Given the value of X_i, the expected, or mean, value of the disturbance term u is zero. That is,

$$E(u \mid X_i) = 0 \qquad \textbf{(3.1)}$$

Recall our discussion in Chapter 2 about the nature of the random term u_i. It represents all those factors that are not specifically introduced in the model.

[1]For further discussion, see Gujarati and Porter, *Basic Econometrics*, 5th ed., McGraw-Hill, New York, 2009.

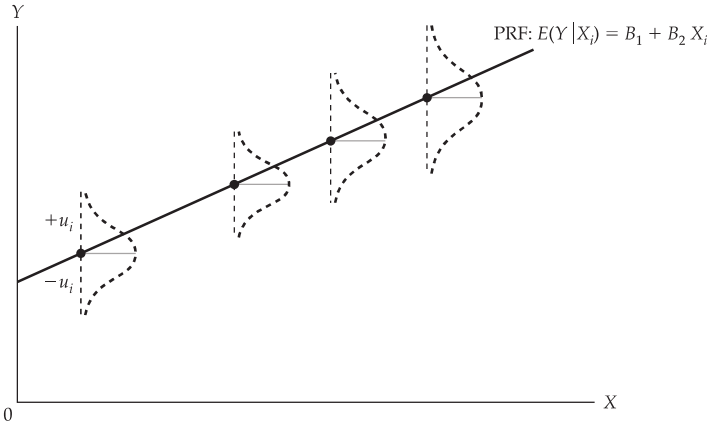

FIGURE 3-1 Conditional distribution of disturbances u_i

What Assumption (3.1) states is that these other factors or forces are not related to X_i (the variable explicitly introduced in the model) and therefore, given the value of X_i, *their mean value is zero.*[2] This is shown in Figure 3-1.

A3.4.

The variance of each u_i is constant, or **homoscedastic** (*homo* means equal and *scedastic* means variance). That is

$$\text{var}(u_i) = \sigma^2 \tag{3.2}$$

Geometrically, this assumption is as shown in Figure 3-2(*a*). This assumption simply means that the conditional distribution of each Y population corresponding to the given value of X has the same variance; that is, the individual Y values are spread around their mean values with the same variance.[3] If this is not the case, then we have **heteroscedasticity,** or **unequal variance,** which is depicted in Figure 3-2(*b*).[4] As this figure shows, the variance of each Y population is different, which is in contrast to Figure 3-2(*a*), where each Y population has the same variance. The CLRM assumes that the variance of u is as shown in Figure 3-2(*a*).

[2]Note that Assumption (3.2) only states that X and u are uncorrelated. Assumption (3.3) adds that not only are X and u uncorrelated, but also that given the value of X, the mean of u (which represents umpteen factors) is zero.

[3]Since the X values are assumed to be given, or nonstochastic, the only source of variation in Y is from u. Therefore, given X_i, the variance of Y_i is the same as that of u_i. In short, the conditional variances of u_i and Y_i are the same, namely, σ^2. Note, however, that the unconditional variance of Y_i, as shown in Appendix B, is $E[Y_i - E(Y)]^2$. As we will see, if the variable X has any impact on Y, the conditional variance of Y will be smaller than the unconditional variance of Y. Incidentally, the sample counterpart of the unconditional variance of Y is $\sum(Y_i - \overline{Y})^2/(n-1)$.

[4]There is a debate in the literature regarding whether it is homoscedasticity or homoskedasticity and heteroscedasticity or heteroskedasticty. Both seem to be acceptable.

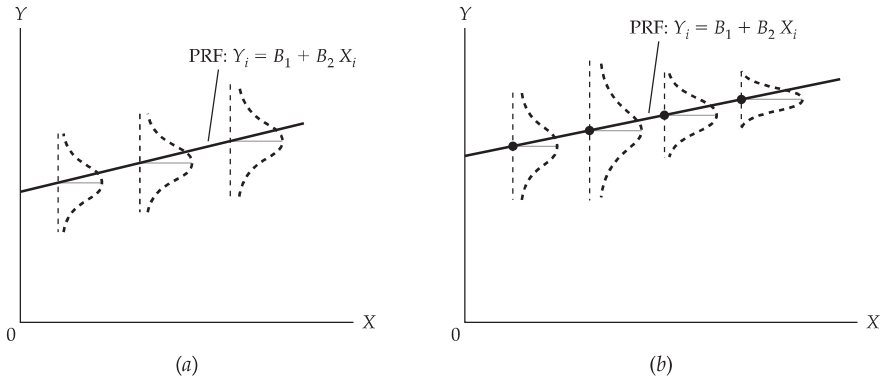

FIGURE 3-2 (*a*) Homoscedasticity (equal variance); (*b*) Heteroscedasticity (unequal variance)

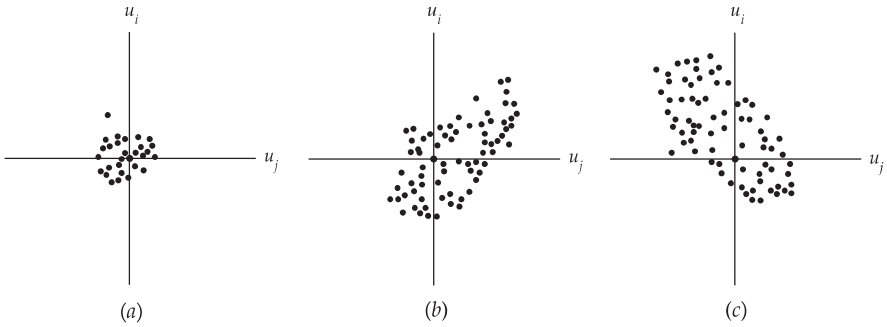

FIGURE 3-3 Patterns of autocorrelation: (*a*) No autocorrelation; (*b*) positive autocorrelation; (*c*) negative autocorrelation

A3.5.

There is no correlation between two error terms. This is the assumption of **no autocorrelation.**

Algebraically, this assumption can be written as

$$\text{cov}\,(u_i, u_j) = 0 \quad i \neq j \tag{3.3}$$

Here *cov* stands for covariance (see Appendix B) and i and j are any two error terms. (*Note:* If $i = j$, Equation (3.3) will give the variance of u, which by Eq. (3.2) is a constant).

Geometrically, Eq. (3.3) can be shown in Figure 3-3.

Assumption (3.5) means that there is no systematic relationship between two error terms. It does not mean that if one u is above the mean value, another error term u will also be above the mean value (for positive correlation), or that if one error term is below the mean value, another error term has to be above the mean value, or vice versa (negative correlation). In short, the assumption of no autocorrelation means the error terms u_i are random.

Since any two error terms are assumed to be uncorrelated, it means that any two Y values will also be uncorrelated; that is, $\text{cov}(Y_i, Y_j) = 0$. This is because $Y_i = B_1 + B_2X_i + u_i$ and given that the B's are fixed numbers and that X is assumed to be fixed, Y will vary as u varies. So, if the u's are uncorrelated, the Y's will be uncorrelated also.

A3.6.

The regression model is correctly specified. Alternatively, there is no *specification bias* or *specification error* in the model used in empirical analysis.

What this assumption implies is that we have included all the variables that affect a particular phenomenon. Thus, if we are studying the demand for automobiles, if we only include prices of automobiles and consumer income and do not take into account variables such as advertising, financing costs, and gasoline prices, we will be committing model specification errors. Of course, it is not easy to determine the "correct" model in any given case, but we will provide some guidelines in Chapter 7.

You might wonder about all these assumptions. Why are they needed? How realistic are they? What happens if they are not true? How do we know that a particular regression model in fact satisfies all these assumptions? Although these questions are certainly pertinent, at this stage of the development of our subject matter, we cannot provide totally satisfactory answers to all of them. However, as we progress through the book, we will see the utility of these assumptions. As a matter of fact, all of Part II is devoted to finding out what happens if one or more of the assumptions of CLRM are not fulfilled.

But keep in mind that in any scientific inquiry we make certain assumptions because they facilitate the development of the subject matter in gradual steps, not because they are necessarily realistic. An analogy might help here. Students of economics are generally introduced to the model of perfect competition before they are introduced to the models of imperfect competition. This is done because the implications derived from this model enable us to better appreciate the models of imperfect competition, not because the model of perfect competition is necessarily realistic, although there are markets that may be reasonably perfectly competitive, such as the stock market or the foreign exchange market.

3.2 VARIANCES AND STANDARD ERRORS OF ORDINARY LEAST SQUARES ESTIMATORS

One immediate result of the assumptions just introduced is that they enable us to estimate the variances and standard errors of the ordinary least squares (OLS) estimators given in Eqs. (2.16) and (2.17). In Appendix D we discuss the basics of estimation theory, including the notions of (point) estimators, their sampling distributions, and the concepts of the variance and standard error of the estimators. Based on our knowledge of those concepts, we know that the

OLS estimators given in Eqs. (2.16) and (2.17) are *random variables,* for their values will change from sample to sample. Naturally, we would like to know something about the sampling variability of these estimators, that is, how they vary from sample to sample. These sampling variabilities, as we know now, are measured by the variances of these estimators, or by their *standard errors* (se), which are the square roots of the variances. The **variances** and **standard errors of the OLS estimators** given in Eqs. (2.16) and (2.17) are as follows:[5]

$$\text{var}\,(b_1) = \sigma^2_{b_1} = \frac{\sum X_i^2}{n\sum x_i^2} \cdot \sigma^2 \tag{3.4}$$

(*Note:* This formula involves both small x and capital X.)

$$\text{se}\,(b_1) = \sqrt{\text{var}\,(b_1)} \tag{3.5}$$

$$\text{var}\,(b_2) = \sigma^2_{b_2} = \frac{\sigma^2}{\sum x_i^2} \tag{3.6}$$

$$\text{se}\,(b_2) = \sqrt{\text{var}\,(b_2)} \tag{3.7}$$

where var = the variance and se = the standard error, and where σ^2 is the variance of the disturbance term u_i, which by the assumption of homoscedasticity is assumed to be the same for each u.

Once σ^2 is known, then all the terms on the right-hand sides of the preceding equations can be easily computed, which will give us the numerical values of the variances and standard errors of the OLS estimators. The homoscedastic σ^2 is estimated from the following formula:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} \tag{3.8}$$

where $\hat{\sigma}^2$ is an estimator of σ^2 (recall we use $\hat{}$ to indicate an estimator) and $\sum e_i^2$ is the **residual sum of squares (RSS),** that is, $\sum (Y_i - \hat{Y}_i)^2$, the sum of the squared difference between the actual Y and the estimated Y. (See the next to the last column of Table 2-4.)

The expression $(n-2)$ is known as the *degrees of freedom* (*d.f.*), which, as noted in Appendix C, is simply the number of independent observations.[6]

Once e_i is computed, as shown in Table 2-4, $\sum e_i^2$ can be computed easily. Incidentally, in passing, note that

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \tag{3.9}$$

[5]The proofs can be found in Gujarati and Porter, *Basic Econometrics,* 5th ed., McGraw-Hill, New York, 2009, pp. 93–94.

[6]Notice that we can compute e_i only when $\hat{Y}_i$ is computed. But to compute the latter, we must first obtain b_1 and b_2. In estimating these two unknowns, we lose 2 d.f. Therefore, although we have n observations, the d.f. are only $(n-2)$.

TABLE 3-1 COMPUTATIONS FOR THE S.A.T. EXAMPLE

Estimator	Formula	Answer	Equation number
$\hat{\sigma}^2$	$\sum\left(\frac{e_i^2}{n-2}\right)$	975.1347	(3.10)
$\hat{\sigma}$	$\sqrt{\hat{\sigma}^2} = \sqrt{975.1347}$	31.2271	(3.11)
var (b_1)	$\left(\frac{\sum X_i^2}{n\sum x_i^2}\right)\sigma^2 = \frac{4.76 \times 10^{10}}{10(1.624 \times 10^{11})}(975.1347)$	285.8153	(3.12)
se (b_1)	$\sqrt{\text{var}(b_1)} = \sqrt{285.8153}$	16.9061	(3.13)
var (b_2)	$\frac{\sigma^2}{\sum x_i^2} = \frac{975.1347}{1.624 \times 10^{11}}$	6.0045×10^{-9}	(3.14)
se (b_2)	$\sqrt{\text{var}(b_2)} = \sqrt{6.0045 \times 10^{-9}}$	0.0000775	(3.15)

Note: The raw data underlying the calculations are given in Table 2-4. In computing the variances of the estimators, σ^2 has been replaced by its estimator, $\hat{\sigma}^2$.

which is known as the **standard error of the regression (SER),** which is simply the standard deviation of the Y values about the estimated regression line.[7] This standard error of regression is often used as a summary measure of the *goodness of fit* of the estimated regression line, a topic discussed in Section 3.6. As you would suspect, the smaller the value of $\hat{\sigma}$, the closer the actual Y value is to its estimated value from the regression model.

Variances and Standard Errors of the Math S.A.T. Score Example

Using the preceding formulas, let us compute the variances and standard errors of our math S.A.T. score example. These calculations are presented in Table 3-1. (See Eqs. [3.10] to [3.15] therein.)

Summary of the Math S.A.T. Score Function

Let us express the estimated S.A.T. score function in the following form:

$$\begin{aligned} \hat{Y}_i &= 432.4138 + 0.0013X_i \\ \text{se} &= (16.9061)(0.000245) \end{aligned} \tag{3.16}$$

where the figures in parentheses are the estimated standard errors. Regression results are sometimes presented in this format (but more on this in Section 3.8). Such a presentation indicates immediately the estimated parameters and their

[7]Note the difference between the standard error of regression $\hat{\sigma}$ and the standard deviation of Y. The latter is measured, as usual, from its mean value, as $S_y = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{n-1}}$, whereas the former is measured from the estimated value (i.e., $\hat{Y}_i$ from the sample regression). See also footnote 3.

standard errors. For example, it tells us that the estimated slope coefficient of the math S.A.T. score function (i.e., the coefficient of the annual family income variable) is 0.0013 and its standard deviation, or standard error, is 0.000245. This is a measure of variability of b_2 from sample to sample.

What use can we make of this finding? Can we say, for example, that our computed b_2 lies within a certain number of standard deviation units from the true B_2? If we can do that, we can state with some confidence (i.e., probability) how good the computed SRF, Equation (3.16), is as an estimate of the true PRF. This is, of course, the topic of hypothesis testing.

But before discussing hypothesis testing, we need a bit more theory. In particular, since b_1 and b_2 are random variables, we must find their **sampling,** or **probability, distributions.** Recall from Appendixes C and D that a random variable (r.v.) has a probability distribution associated with it. Once we determine the sampling distributions of our two estimators, as we will show in Section 3.4, the task of hypothesis testing becomes straightforward. But even before that we answer an important question: Why do we use the OLS method?

3.3 WHY OLS? THE PROPERTIES OF OLS ESTIMATORS

The method of OLS is used popularly not only because it is easy to use but also because it has some strong theoretical properties, which are summarized in the well-known **Gauss-Markov theorem.**

Gauss-Markov Theorem

Given the assumptions of the classical linear regression model, the OLS estimators have minimum variance in the class of linear estimators; that is, they are **BLUE** (best linear unbiased estimators).

We provide an overview of the **BLUE property** in Appendix D. In short, the OLS estimators have the following properties:[8]

1. b_1 and b_2 are linear estimators; that is, they are linear functions of the random variable Y, which is evident from Equations (2.16) and (2.17).
2. They are unbiased; that is, $E(b_1) = B_1$ and $E(b_2) = B_2$. Therefore, in repeated applications, on average, b_1 and b_2 will coincide with their true values B_1 and B_2, respectively.
3. $E(\hat{\sigma}^2) = \sigma^2$ that is, the OLS estimator of the error variance is unbiased. In repeated applications, on average, the estimated value of the error variance will converge to its true value.
4. b_1 and b_2 are *efficient* estimators; that is, var (b_1) is less than the variance of any other linear unbiased estimator of B_1, and var (b_2) is less than the

[8]For proof, see Gujarati and Porter, *Basic Econometrics,* 5th ed., McGraw-Hill, New York, 2009, pp. 95–96.

variance of any other linear unbiased estimator of B_2. Therefore, we will be able to estimate the true B_1 and B_2 more precisely if we use OLS rather than any other method that also gives linear unbiased estimators of the true parameters.

The upshot of the preceding discussion is that the OLS estimators possess many desirable statistical properties that we discuss in Appendix D. It is for this reason that the OLS method has been used popularly in regression analysis, as well as for its intuitive appeal and ease of use.

Monte Carlo Experiment

In theory the OLS estimators are unbiased, but how do we know that in practice this is the case? To find out, let us conduct the following Monte Carlo experiment.

Assume that we are given the following information:

$$\begin{aligned} Y_i &= B_1 + B_2 X_i + u_i \\ &= 1.5 + 2.0 X_i + u_i \end{aligned}$$

where $u_i \sim N(0, 4)$.

That is, we are told that the true values of the intercept and slope coefficients are 1.5 and 2.0, respectively, and that the error term follows the normal distribution with a mean of zero and a variance of 4. Now suppose you are given 10 values of X: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

Given this information, you can proceed as follows. Using any statistical package, you generate 10 values of u_i from a normal distribution with mean zero and variance 4. Given B_1, B_2, the 10 values of X, and the 10 values of u_i generated from the normal distribution, you will then obtain 10 values of Y from the preceding equation. Call this experiment or sample number 1. Go to the normal distribution table, collect another 10 values of u_i, generate another 10 values of Y, and call it sample number 2. In this manner obtain 21 samples.

For each sample of 10 values, regress Y_i generated above on the X values and obtain b_1, b_2, and $\hat{\sigma}^2$. Repeat this exercise for all 21 samples. Therefore, you will have 21 values each of b_1, b_2, and $\hat{\sigma}^2$. We conducted this experiment and obtained the results shown in Table 3-2.

From the data given in this table, we have computed the mean, or average, values of b_1, b_2, and $\hat{\sigma}^2$, which are, respectively, 1.4526, 1.9665, and 4.4743, whereas the true values of the corresponding coefficients, as we know, are 1.5, 2.0, and 4.0.

What conclusion can we draw from this experiment? It seems that if we apply the method of least squares time and again, *on average,* the values of the estimated parameters will be equal to their true (population parameter) values. That is, OLS estimators are unbiased. In the present example, had we conducted more than 21 sampling experiments, we would have come much closer to the true values.

TABLE 3-2 MONTE CARLO EXPERIMENT: $Y_i = 1.5 + 2X_i + u_i$; $u \sim N(0, 4)$

b_1	b_2	$\hat{\sigma}^2$
2.247	1.840	2.7159
0.360	2.090	7.1663
−2.483	2.558	3.3306
0.220	2.180	2.0794
3.070	1.620	4.3932
2.570	1.830	7.1770
2.551	1.928	5.7552
0.060	2.070	3.6176
−2.170	2.537	3.4708
1.470	2.020	4.4479
2.540	1.970	2.1756
2.340	1.960	2.8291
0.775	2.050	1.5252
3.020	1.740	1.5104
0.810	1.940	4.7830
1.890	1.890	7.3658
2.760	1.820	1.8036
−0.136	2.130	1.8796
0.950	2.030	4.9908
2.960	1.840	4.5514
3.430	1.740	5.2258
$\bar{b}_1 = 1.4526$	$\bar{b}_2 = 1.9665$	$\bar{\hat{\sigma}}^2 = 4.4743$

3.4 THE SAMPLING, OR PROBABILITY, DISTRIBUTIONS OF OLS ESTIMATORS

Now that we have seen how to compute the OLS estimators and their standard errors and have examined some of the properties of these estimators, we need to find the sampling distributions of these estimators. Without that knowledge we will not be able to engage in hypothesis testing. The general notion of sampling distribution of an estimator is discussed in Appendix C (see Section C.2).

To derive the sampling distributions of the OLS estimators b_1 and b_2, we need to add one more assumption to the list of assumptions of the CLRM. This assumption is

A3.7.

In the PRF $Y_i = B_1 + B_2X_i + u_i$ the error term u_i follows the *normal distribution* with mean zero and variance σ^2. That is,

$$u_i \sim N(0, \sigma^2) \tag{3.17}$$

What is the rationale for this assumption? There is a celebrated theorem in statistics, known as the **central limit theorem (CLT),** which we discuss in Appendix C (see Section C.1), which states that:

Central Limit Theorem

If there is a large number of independent and identically distributed random variables, then, with a few exceptions,[9] the distribution of their sum tends to be a normal distribution as the number of such variables increases indefinitely.

Recall from Chapter 2 our discussion about the nature of the error term, u_i. As shown in Section 2.4, the error term represents the influence of all those forces that affect Y but are not specifically included in the regression model because there are so many of them and the individual effect of any one such force (i.e., variable) on Y may be too minor. If all these forces are random, and if we let u represent the sum of all these forces, then by invoking the CLT we can assume that the error term u follows the normal distribution. We have already assumed that the mean value of u_i is zero and that its variance, following the homoscedasticity assumption, is the constant σ^2. Hence, we have Equation (3.17).

But how does the assumption that u follows the normal distribution help us to find out the probability distributions of b_1 and b_2? Here we make use of another property of the normal distribution discussed in Appendix C, namely, *any linear function of a normally distributed variable is itself normally distributed.* Does this mean that if we prove that b_1 and b_2 are linear functions of the normally distributed variable u_i, they themselves are normally distributed? That's right! You can indeed prove that these two OLS estimators are in fact linear functions of the normally distributed u_i. (For proof, see Exercise 3.24).[10]

Now we know from Appendix C that a normally distributed r.v. has two parameters, the mean and the variance. What are the parameters of the normally distributed b_1 and b_2? They are as follows:

$$b_1 \sim N\left(B_1, \sigma^2_{b_1}\right) \tag{3.18}$$

$$b_2 \sim N\left(B_2, \sigma^2_{b_2}\right) \tag{3.19}$$

where the variances of b_1 and b_2 are as given in Eq. (3.4) and Eq. (3.6).

In short, b_1 and b_2 each follow the normal distribution with their means equal to true B_1 and B_2 and their variances given by Eqs. (3.4) and (3.6) developed previously. Geometrically, the distributions of these estimators are as shown in Figure 3-4.

[9]One exception is the Cauchy probability distribution, which has no mean or variance.

[10]It may also be noted that since $Y_i = B_1 + B_2X_i + u_i$ if $u_i \sim N(0, \sigma^2)$, then $Y_i \sim N(B_1 + B_2X_i, \sigma^2)$ because Y_i is a linear combination of u_i. (Note that B_1, B_2 are constants and X_i fixed).

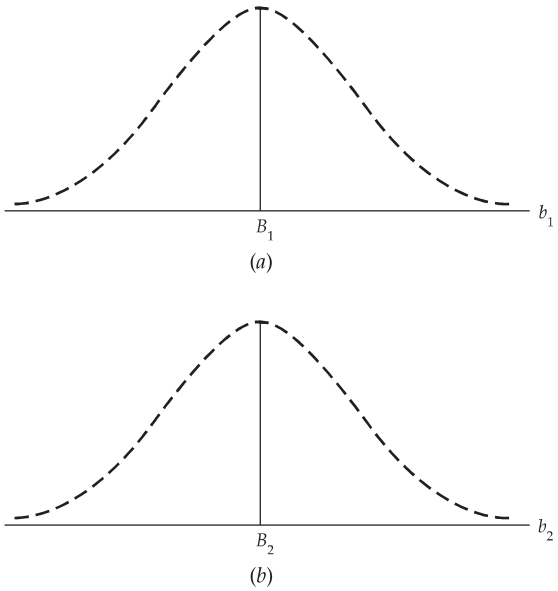

FIGURE 3-4 (Normal) sampling distributions of b_1 and b_2

3.5 HYPOTHESIS TESTING

Recall that estimation and hypothesis testing are the two main branches of statistical inference. In Chapter 2 we showed how OLS helps us to estimate the parameters of linear regression models. In this chapter the classical framework enabled us to examine some of the properties of OLS estimators. With the added assumption that the error term u_i is normally distributed, we were able to find the sampling (or probability) distributions of the OLS estimators, namely, the normal distribution. With this knowledge we are now equipped to deal with the topic of hypothesis testing in the context of regression analysis.

Let us return to our math S.A.T. example. The estimated math S.A.T. score function is given in Eq. (2.20). Suppose someone suggests that annual family income has no relationship to a student's math S.A.T. score.

$$H_0: B_2 = 0$$

In applied regression analysis such a **"zero" null hypothesis,** the so-called **straw man hypothesis,** is deliberately chosen to find out whether Y is related to X at all. If there is no relationship between Y and X to begin with, then testing a hypothesis that $B_2 = -2$ or any other value is meaningless. Of course, if the zero null hypothesis is sustainable, there is no point at all in including X in the model. Therefore, if X really belongs in the model, you would fully expect to reject the zero null hypothesis H_0 in favor of the *alternative hypothesis* H_1, which says, for example, that $B_2 \neq 0$; that is, the slope coefficient is different from zero. It could be positive or it could be negative.

Our numerical results show that $b_2 = 0.0013$. You would therefore expect that the zero null hypothesis is not tenable in this case. But we cannot look at the numerical results alone, for we know that because of sampling fluctuations, the numerical value will change from sample to sample. Obviously, we need some formal testing procedure to reject or not reject the null hypothesis. How do we proceed?

This should not be a problem now, for in Equation (3.19) we have shown that b_2 follows the *normal distribution* with mean $= B_2$ and $\text{var}(b_2) = \sigma^2 / \sum x_i^2$. Then, following our discussion about hypothesis testing in Appendix D, Section D.5, we can use either:

1. The confidence interval approach or
2. The test of significance approach

to test any hypotheses about B_2 as well as B_1.

Since b_2 follows the normal distribution, with the mean and the variance stated in expression (3.19), we know that

$$\begin{aligned} Z &= \frac{b_2 - B_2}{\text{se}(b_2)} \\ &= \frac{b_2 - B_2}{\sigma \Big/ \sqrt{\sum x_i^2}} \sim N(0, 1) \end{aligned} \tag{3.20}$$

follows the *standard normal distribution.* From Appendix C we know the properties of the standard normal distribution, particularly, the property that ≈95 percent of the area of the normal distribution lies within two standard deviation units of the mean value, where ≈ means approximately. Therefore, if our null hypothesis is $B_2 = 0$ and the computed $b_2 = 0.0013$, we can find out the probability of obtaining such a value from the Z, or standard normal, distribution (Appendix E, Table E-1). If this probability is very small, we can reject the null hypothesis, but if it is large, say, greater than 10 percent, we may not reject the null hypothesis. All this is familiar material from Appendixes C and D.

But, there is a hitch! To use Equation (3.20) we must know the true σ^2. This is not known, but we can estimate it by using $\hat{\sigma}^2$ given in Eq. (3.8). However, if we replace σ in Eq. (3.20) by its estimator $\hat{\sigma}$, then, as shown in Appendix C, Eq. (C.8), the right-hand side of Eq. (3.20) follows the *t distribution* with $(n - 2)$ d.f., not the standard normal distribution; that is,

$$\frac{b_2 - B_2}{\hat{\sigma} \Big/ \sqrt{\sum x_i^2}} \sim t_{n-2} \tag{3.21}$$

Or, more generally,

$$\frac{b_2 - B_2}{\text{se}(b_2)} \sim t_{n-2} \tag{3.22}$$

Note that we lose 2 d.f. in computing $\hat{\sigma}^2$ for reasons stated earlier.

Therefore, to test the null hypothesis in the present case, we have to use the t distribution in lieu of the (standard) normal distribution. But the procedure of hypothesis testing remains the same, as explained in Appendix D.

Testing $H_0: B_2 = 0$ versus $H_1: B_2 \neq 0$: The Confidence Interval Approach

For our math S.A.T. example we have 10 observations, hence the d.f. are $(10 - 2) = 8$. Let us assume that α, the level of significance or the probability of committing a type I error, is fixed at 5 percent. Since the alternative hypothesis is two-sided, from the t table given in Appendix E, Table E-2, we find that for 8 d.f.,

$$P(-2.306 \leq t \leq 2.306) = 0.95 \tag{3.23}$$

That is, the probability that a t value (for 8 d.f.) lies between the limits (−2.306, 2.306) is 0.95 or 95 percent; these, as we know, are the *critical* t values. Now by substituting for t from expression (3.21) into the preceding equation, we obtain

$$P\left(-2.306 \leq \frac{b_2 - B_2}{\hat{\sigma}\Big/\sqrt{\sum x_i^2}} \leq 2.306\right) = 0.95 \tag{3.24}$$

Rearranging inequality (3.24), we obtain

$$P\left(b_2 - 2.306\frac{\hat{\sigma}}{\sqrt{\sum x_i^2}} \leq B_2 \leq b_2 + 2.306\frac{\hat{\sigma}}{\sqrt{\sum x_i^2}}\right) = 0.95 \tag{3.25}$$

Or, more generally,

$$P[(b_2 - 2.306\,\text{se}(b_2) \leq B_2 \leq b_2 + 2.306\,\text{se}(b_2)] = 0.95 \tag{3.26}$$

which provides *a 95% confidence interval for* B_2. In repeated applications 95 out of 100 such intervals will include the true B_2. As noted previously, in the language of hypothesis testing such a confidence interval is known as the *region of acceptance* (of H_0) and the area outside the confidence interval is known as the *rejection region* (of H_0).

Geometrically, the 95% confidence interval is shown in Figure 3-5(*a*).

Now following our discussion in Appendix D, if this interval (i.e., the acceptance region) includes the null-hypothesized value of B_2, we do not reject the hypothesis. But if it lies outside the confidence interval (i.e., it lies in the rejection region), we reject the null hypothesis, bearing in mind that in making either of these decisions we are taking a chance of being wrong a certain percent, say, 5 percent, of the time.

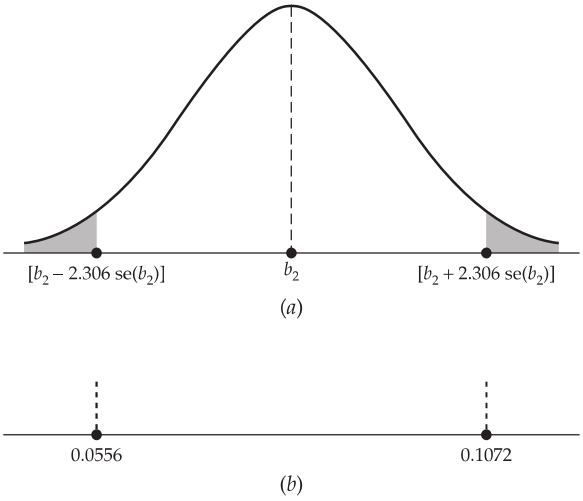

FIGURE 3-5 (*a*) 95% confidence interval for B_2 (8 d.f.); (*b*) 95% confidence interval for the slope coefficient of the math S.A.T. score example

All that remains to be done for our math S.A.T. score example is to obtain the numerical value of this interval. But that is now easy, for we have already obtained se(b_2) = 0.000245, as shown in Eq. (3.16). Substituting this value in Eq. (3.26), we now obtain the 95% confidence interval as shown in Figure 3-5(*b*).

$$0.0013 - 2.306(0.000245) \leq B_2 \leq 0.0013 + 2.306(0.000245)$$

That is,

$$0.00074 \leq B_2 \leq 0.00187 \qquad \textbf{(3.27)}$$

Since this interval does not include the null-hypothesized value of 0, we can reject the null hypothesis that annual family income is not related to math S.A.T. scores. Put positively, income does have a relationship to math S.A.T. scores.

A cautionary note: As noted in Appendix D, although the statement given in Eq. (3.26) is true, we *cannot say* that the probability is 95 percent that the particular interval in Eq. (3.27) includes the true B_2, for unlike Eq. (3.26), expression (3.27) is not a random interval; it is fixed. Therefore, the probability is either 1 or 0 that the interval in Eq. (3.27) includes B_2. We can only say that if we construct 100 intervals like the interval in Eq. (3.27), 95 out of 100 such intervals will include the true B_2; we cannot guarantee that this particular interval will necessarily include B_2.

Following a similar procedure exactly, the reader should verify that the 95% confidence interval for the intercept term B_1 is

$$393.4283 \leq B_1 \leq 471.3993 \qquad \textbf{(3.28)}$$

If, for example, H_0:$B_1 = 0$ vs. H_1:$B_1 \neq 0$, obviously this null hypothesis will be rejected too, for the preceding 95% confidence interval does not include 0.

On the other hand, if the null hypothesis were that the true intercept term is 400, we would not reject this null hypothesis because the 95% confidence interval includes this value.

The Test of Significance Approach to Hypothesis Testing

The key idea underlying this approach to hypothesis testing is that of a *test statistic* (see Appendix D) and the *sampling distribution* of the test statistic under the null hypothesis, H_0. The decision to accept or reject H_0 is made on the basis of the value of the test statistic obtained from the sample data.

To illustrate this approach, recall that

$$t = \frac{b_2 - B_2}{\text{se}(b_2)} \tag{3.22}$$

follows the t distribution with $(n - 2)$ d.f. Now if we let

$$H_0: B_2 = B_2^*$$

where B_2^* is a *specific numerical value* of B_2 (e.g., $B_2^* = 0$), then

$$\begin{aligned} t &= \frac{b_2 - B_2^*}{\text{se}(b_2)} \\ &= \frac{\text{estimator} - \text{hypothesized value}}{\text{standard error of the estimator}} \end{aligned} \tag{3.29}$$

can be readily computed from the sample data. Since all the quantities in Equation (3.29) are now known, we can use the t value computed from Eq. (3.29) as the test statistic, which follows the t distribution with $(n - 2)$ d.f. Appropriately, the testing procedure is called the ***t* test.**[11]

Now to use the t test in any concrete application, we need to know three things:

1. The d.f., which are always $(n - 2)$ for the two-variable model
2. The level of significance, α, which is a matter of personal choice, although 1, 5, or 10 percent levels are usually used in empirical analysis. Instead of arbitrarily choosing the α value, you can find the *p value* (the exact level of significance as described in Appendix D) and reject the null hypothesis if the computed *p value* is sufficiently low.
3. Whether we use a one-tailed or two-tailed test (see Table D-2 and Figure D-7).

[11]The difference between the confidence interval and the test of significance approaches lies in the fact that in the former we do not know what the true B_2 is and therefore try to guess it by establishing a $(1 - \alpha)$ confidence interval. In the test of significance approach, on the other hand, we hypothesize what the true B_2 $(= B_2^*)$ is and try to find out if the sample value b_2 is sufficiently close to (the hypothesized) B_2^*.

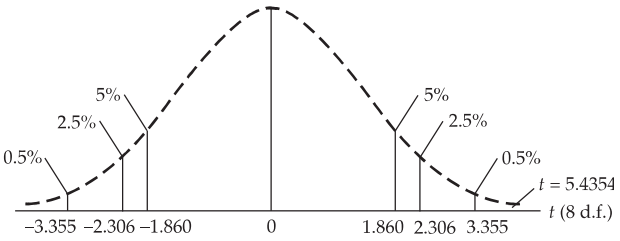

FIGURE 3-6 The *t* distribution for 8 d.f.

Math S.A.T. Example Continued

1. **A Two-Tailed Test** Assume that H_0:$B_2 = 0$ and H_1:$B_2 \neq 0$. Using Eq. (3.29), we find that

$$t = \frac{0.0013}{0.000245} = 5.4354 \tag{3.30}$$

Now from the *t* table given in Appendix E, Table E-2, we find that for 8 d.f. we have the following critical *t* values (two-tailed) (see Figure 3-6):

Level of significance	Critical *t*
0.01	3.355
0.05	2.306
0.10	1.860

In Appendix D, Table D-2 we stated that, in the case of the two-tailed *t* test, if the computed $|t|$, the absolute value of *t*, exceeds the critical *t* value at the chosen level of significance, we can reject the null hypothesis. Therefore, in the present case we can reject the null hypothesis that the true B_2 (i.e., the income coefficient) is zero because the computed $|t|$ of 5.4354 far exceeds the critical *t* value even at the 1% level of significance. We reached the same conclusion on the basis of the confidence interval shown in Eq. (3.27), which should not be surprising because *the confidence interval and the test of significance approaches to hypothesis testing are merely two sides of the same coin.*

Incidentally, in the present example the *p value* (i.e., probability value) of the *t* statistic of 5.4354 is about 0.0006. Thus, if we were to reject the null hypothesis that the true slope coefficient is zero at this *p value*, we would be wrong in six out of ten thousand occasions.

2. **A One-Tailed Test** Since the income coefficient in the math S.A.T. score function is expected to be positive, a realistic set of hypotheses would be H_0:$B_2 \leq 0$ and H_1:$B_2 > 0$; here the alternative hypothesis is one-sided.

The *t*-testing procedure remains exactly the same as before, except, as noted in Appendix D, Table D-2, the probability of committing a type I error is *not* divided equally between the two tails of the *t* distribution but is concentrated in only one tail, either left or right. In the present case it will be the right tail. (Why?) For 8 d.f. we observe from the *t* table (Appendix E, Table E-2) that the critical *t* value (right-tailed) is

Level of significance	Critical *t*
0.01	2.896
0.05	1.860
0.10	1.397

For the math S.A.T. example, we first compute the *t* value as if the null hypothesis were that $B_2 = 0$. We have already seen that this *t* value is

$$t = 5.4354 \qquad \textbf{(3.30)}$$

Since this *t* value exceeds any of the critical values shown in the preceding table, following the rules laid down in Appendix D, Table D-2, we can reject the hypothesis that annual family income has no relationship to math S.A.T. scores; actually it has a positive effect (i.e., $B_2 > 0$) (see Figure 3-7).

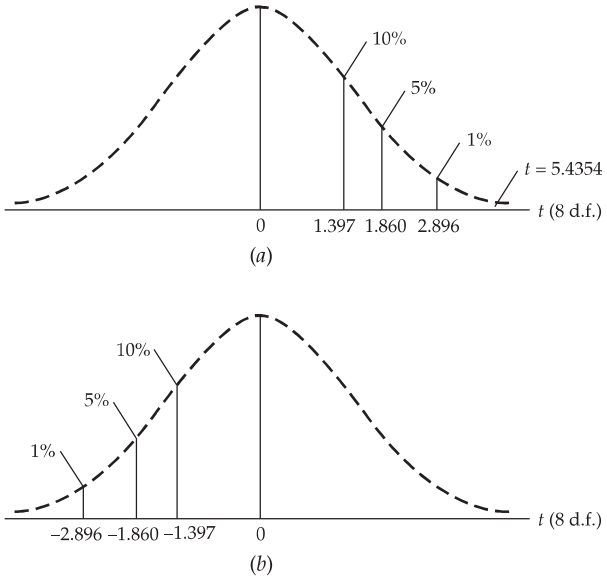

FIGURE 3-7 One-tailed *t* test: (*a*) Right-tailed; (*b*) left-tailed

3.6 HOW GOOD IS THE FITTED REGRESSION LINE: THE COEFFICIENT OF DETERMINATION, r^2

Our finding in the preceding section that on the basis of the t test both the estimated intercept and slope coefficients are *individually* statistically significant (i.e., significantly different from zero) suggests that the SRF, Eq. (3.16), shown in Figure 2-6 seems to "fit" the data "reasonably" well. Of course, not each actual Y value lies on the estimated PRF. That is, not all $e_i = (Y_i - \hat{Y}_i)$ are zero; as Table 2-4 shows, some e are positive and some are negative. Can we develop an overall measure of "goodness of fit" that will tell us how well the estimated regression line, Eq. (3.16), fits the actual Y values? Indeed, such a measure has been developed and is known as the **coefficient of determination,** denoted by the symbol r^2 (read as r squared). To see how r^2 is computed, we proceed as follows.

Recall that

$$Y_i = \hat{Y}_i + e_i \qquad \textbf{(Eq. 2.6)}$$

Let us express this equation in a slightly different but equivalent form (see Figure 3-8) as

$$\underset{\substack{\text{Variation in } Y_i \\ \text{from its mean value}}}{(Y_i - \overline{Y})} = \underset{\substack{\text{Variation in } Y_i \text{ explained} \\ \text{by } X(=\hat{Y}_i) \text{ around} \\ \text{its mean value} \\ (\text{Note: } \overline{Y} = \overline{\hat{Y}})}}{(\hat{Y}_i - \overline{Y})} + \underset{\substack{\text{Unexplained or} \\ \text{residual variation}}}{(Y_i - \hat{Y}_i)(\text{i.e., } e_i)} \qquad \textbf{(3.31)}$$

Now, letting small letters indicate deviations from mean values, we can write the preceding equation as

$$y_i = \hat{y}_i + e_i \qquad \textbf{(3.32)}$$

(*Note:* $y_i = (Y_i - \overline{Y})$, etc.) Also, note that $\bar{e} = 0$, as a result of which $\overline{Y} = \overline{\hat{Y}}$; that is, the mean values of the actual Y and the estimated Y are the same. Or

$$y_i = b_2 x_i + e_i \qquad \textbf{(3.33)}$$

since $\hat{y}_i = b_2 x_i$.

Now squaring Equation (3.33) on both sides and summing over the sample, we obtain, after simple algebraic manipulation,

$$\sum y_i^2 = \sum \hat{y}_i^{\,2} + \sum e_i^2 \qquad \textbf{(3.34)}$$

Or, equivalently,

$$\sum y_i^2 = b_2^2 \sum x_i^2 + \sum e_i^2 \qquad \textbf{(3.35)}$$

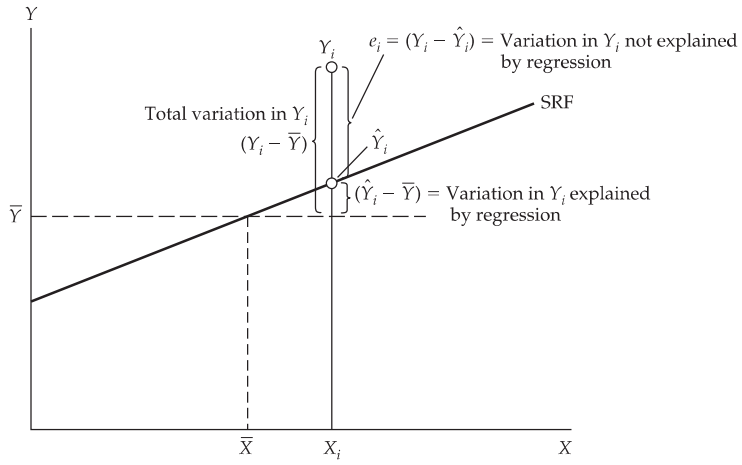

FIGURE 3-8 Breakdown of total variation in Y_i

This is an important relationship, as we will see. For proof of Equation (3.35), see Problem 3.25.

The various sums of squares appearing in Eq. (3.35) can be defined as follows:

$\sum y_i^2$ = the total variation[12] of the actual Y values about their sample mean $\overline{Y}$, which may be called the **total sum of squares (TSS).**

$\sum \hat{y}_i^2$ = the total variation of the estimated Y values about their mean value ($\hat{Y} = \overline{Y}$), which may be called appropriately the *sum of squares due to regression* (i.e., due to the explanatory variable [s]), or simply the **explained sum of squares (ESS).**

$\sum e_i^2$ = as before, the residual sum of squares (RSS) or residual or unexplained variation of the Y values about the regression line.

Put simply, then, Eq. (3.35) is

$$\text{TSS} = \text{ESS} + \text{RSS} \tag{3.36}$$

and shows that the total variation in the observed Y values about their mean value can be partitioned into two parts, one attributable to the regression line and the other to random forces, because not all actual Y observations lie on the fitted line. All this can be seen clearly from Figure 3-8 (see also Fig. 2-6).

Now if the chosen SRF fits the data quite well, ESS should be much larger than RSS. If all actual Y lie on the fitted SRF, ESS will be equal to TSS, and RSS will be zero. On the other hand, if the SRF fits the data poorly, RSS will be much larger than ESS. In the extreme, if X explains no variation at all in Y, ESS will be zero and RSS will equal TSS. These are, however, polar cases. Typically, neither

[12]The terms *variation* and *variance* are different. *Variation* means the sum of squares of deviations of a variable from its mean value. *Variance* is this sum divided by the appropriate d.f. In short, variance = variation/d.f.

ESS nor RSS will be zero. If ESS is relatively larger than RSS, the SRF will explain a substantial proportion of the variation in Y. If RSS is relatively larger than ESS, the SRF will explain only some part of the variation of Y. All these qualitative statements are intuitively easy to understand and can be readily quantified. If we divide Equation (3.36) by TSS on both sides, we obtain

$$1 = \frac{\text{ESS}}{\text{TSS}} + \frac{\text{RSS}}{\text{TSS}} \tag{3.37}$$

Now let us define

$$r^2 = \frac{\text{ESS}}{\text{TSS}} \tag{3.38}$$

The quantity r^2 thus defined is known as the (sample) coefficient of determination and is the most commonly used measure of the goodness of fit of a regression line. Verbally, r^2 *measures the proportion or percentage of the total variation in Y explained by the regression model.*

Two properties of r^2 may be noted:

1. It is a non-negative quantity. (Why?)
2. Its limits are $0 \leq r^2 \leq 1$ since a part (ESS) cannot be greater than the whole (TSS).[13] An r^2 of 1 means a "perfect fit," for the entire variation in Y is explained by the regression. An r^2 of zero means no relationship between Y and X whatsoever.

Formulas to Compute r^2

Using Equation (3.38), Equation (3.37) can be written as

$$\begin{aligned} 1 &= r^2 + \frac{\text{RSS}}{\text{TSS}} \\ &= r^2 + \frac{\sum e_i^2}{\sum y_i^2} \end{aligned} \tag{3.39}$$

Therefore,

$$r^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2} \tag{3.40}$$

There are several equivalent formulas to compute r^2, which are given in Question 3.5.

[13]This statement assumes that an intercept term is included in the regression model. More on this in Chapter 5.

r^2 for the Math S.A.T. Example

From the data given in Table 2-4, and using formula (3.40), we obtain the following r^2 value for our math S.A.T. score example:

$$\begin{aligned} r^2 &= 1 - \frac{7801.0776}{36610} \\ &= 0.7869 \end{aligned} \tag{3.41}$$

Since r^2 can at most be 1, the computed r^2 is pretty high. In our math S.A.T. example X, the income variable, explains about 79 percent of the variation in math S.A.T. scores. In this case we can say that the sample regression (3.16) gives an excellent fit.

It may be noted that $(1 - r^2)$, the proportion of variation in Y not explained by X, is called, perhaps appropriately, the **coefficient of alienation.**

The Coefficient of Correlation, r

In Appendix B, we introduce the sample **coefficient of correlation, *r,*** as a measure of the strength of the linear relationship between two variables Y and X and show that r can be computed from formula (B.46), which can also be written as

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{(X_i - \overline{X})^2(Y_i - \overline{Y})^2}} \tag{3.42}$$

$$= \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \tag{3.43}$$

But this coefficient of correlation can also be computed from the coefficient of determination, r^2, as follows:

$$r = \pm\sqrt{r^2} \tag{3.44}$$

Since most regression computer packages routinely compute r^2, r can be computed easily. The only question is about the sign of r. However, that can be determined easily from the nature of the problem. In our math S.A.T. example, since math S.A.T. scores and annual family income are expected to be positively related, the r value in this case will be positive. In general, though, r has the same sign as the slope coefficient, which should be clear from formulas (2.17) and (3.43).

Thus, for the math S.A.T. example,

$$r = \sqrt{0.7869} = 0.8871 \tag{3.45}$$

In our example, math S.A.T. scores and annual family income are highly positively correlated, a finding that is not surprising.

Incidentally, if you use formula (3.43) to compute r between the actual Y values in the sample and the estimated Y_i values ($= \hat{Y}_i$) from the given model, and square this r value, the squared r is precisely equal to the r^2 value obtained from Eq. (3.42). For proof, see Question 3.5. You can verify this from the data given in Table 2-4. As you would expect, the closer the estimated Y values are to the actual Y values in the sample, the higher the r^2 value will be.

3.7 REPORTING THE RESULTS OF REGRESSION ANALYSIS

There are various ways of reporting the results of regression analysis. Until the advent of statistical software, regression results were presented in the format shown in Equation (3.46). Many journal articles still present regression results in this format. For our math S.A.T. score example, we have:

$$\begin{aligned} \hat{Y}_i &= 432.4138 + 0.0013X_i \\ \text{se} &= (16.9061)(0.000245) \\ t &= (25.5774)(0.0006) \qquad r^2 = 0.7849 \\ p \text{ value} &= (5.85 \times 10^{-9})(0.0006) \qquad \text{d.f.} = 8 \end{aligned} \tag{3.46}$$

In Equation (3.46) the figures in the first set of parentheses are the estimated standard errors (se) of the estimated regression coefficients. Those in the second set of parentheses are the estimated t values computed from Eq. (3.22) under the null hypothesis that the true population value of each regression coefficient individually is zero (i.e., the t values given are simply the ratios of the estimated coefficients to their standard errors). And those in the third set of parentheses are the p values of the computed t values.[14] As a matter of convention, from now on, if we do not specify a specific null hypothesis, then we will assume that it is the *zero null hypothesis* (i.e., the population parameter assumes zero value). And if we reject it (i.e., when the test statistic is significant), it means that the true population value is different from zero.

One advantage of reporting the regression results in the preceding format is that we can see at once whether each estimated coefficient is individually statistically significant, that is, significantly different from zero. By quoting the p values we can determine the exact level of significance of the estimated t value. Thus the t value of the estimated slope coefficient is 5.4354, whose p value is practically zero. As we note in Appendix D, *the lower the p value, the greater the evidence against the null hypothesis.*

A warning is in order here. When deciding whether to reject or not reject a null hypothesis, determine *beforehand* what level of the p value (call it the critical p value) you are willing to accept and then compare the computed p value with the critical p value. If the computed p value is smaller than the critical p value, the null hypothesis can be rejected. But if it is greater than the critical

[14]The t table in Appendix E of this book (Table E-2) can now be replaced by electronic tables that will compute the p values to several digits. This is also true of the normal, chi-square, and the F tables (Appendix E, Tables E-4 and E-3, respectively).

p value the null hypothesis may not be rejected. If you feel comfortable with the tradition of fixing the critical p value at the conventional 1, 5, or 10 percent level, that is fine. In Eq. (3.46), the actual p value (i.e., the exact level of significance) of the t coefficient of 5.4354 is 0.0006. If we had chosen the critical p value at 5 percent, obviously we would reject the null hypothesis, for the computed p value of 0.0006 is much smaller than 5 percent.

Of course, any null hypothesis (besides the zero null hypothesis) can be tested easily by making use of the t test discussed earlier. Thus, if the null hypothesis is that the true intercept term is 450 and if H_1: $B_1 \neq 450$, the t value will be

$$t = \frac{432.4138 - 450}{16.9061} = -1.0402$$

The p value of obtaining such a t value is about 0.3287, which is obtained from electronic tables. If you had fixed the critical p value at the 10 percent level, you would not reject the null hypothesis, for the computed p value is much greater than the critical p value.

The zero null hypothesis, as mentioned before, is essentially a kind of straw man. It is usually adopted for strategic reasons—to "dramatize" the statistical significance (i.e., importance) of an estimated coefficient.

3.8 COMPUTER OUTPUT OF THE MATH S.A.T. SCORE EXAMPLE

Since these days we rarely run regressions manually, it may be useful to produce the actual output of regression analysis obtained from a statistical software package. Below we give the selected output of our math S.A.T. example obtained from EViews.

Dependent Variable: Y
Method: Least Squares
Sample: 1 10
Included observations: 10

	Coefficient	Std. Error	t-Statistic	Prob.
C	432.4138	16.90607	25.57742	0.0000
X	0.001332	0.000245	5.435396	0.0006
R-squared		0.786914		
S.E. of regression		31.22715		
Sum squared resid		7801.078		

In this output, C denotes the constant term (i.e., intercept); Prob. is the p value; sum of squared resid is the RSS ($= \sum e_i^2$); and S.E. of regression is the standard error of the regression. The t values presented in this table are computed under the (null) hypothesis that the corresponding population regression coefficients are zero.

We also show (in Figure 3-9) how EViews presents the actual and estimated Y values as well as the residuals (i.e., e_i) in graphic form:

Actual Y_i	Fitted $\hat{Y}_i$	Residual e_i	Residual Plot (−) (0) (+)
410.000	439.073	−29.0733	
420.000	452.392	−32.3922	
440.000	465.711	−25.7112	
490.000	479.030	10.9698	
530.000	492.349	37.6509	
530.000	505.668	24.3319	
550.000	518.987	31.0129	
540.000	532.306	07.69397	
570.000	552.284	17.7155	
590.000	632.198	−42.1983	

FIGURE 3-9 Actual and fitted Y values and residuals for the math S.A.T. example

3.9 NORMALITY TESTS

Before we leave our math S.A.T. example, we need to look at the regression results given in Eq. (3.46). Remember that our statistical testing procedure is based on the assumption that the error term u_i is normally distributed. How do we find out if this is the case in our example, since we do not directly observe the true errors u_i? We have the residuals, e_i, which are proxies for u_i. Therefore, we will have to use the e_i to learn something about the normality of u_i. There are several tests of normality, but here we will consider only three comparatively simple tests.[15]

Histograms of Residuals

A histogram of residuals is a simple graphical device that is used to learn something about the shape of the probability density function (PDF) of a random variable. On the horizontal axis, we divide the values of the variable of interest (e.g., OLS residuals) into suitable intervals, and in each class interval, we erect rectangles equal in height to the number of observations (i.e., frequency) in that class interval.

If you mentally superimpose the bell-shaped normal distribution curve on this histogram, you might get some idea about the nature of the probability distribution of the variable of interest.

It is always a good practice to plot the histogram of residuals from any regression to get some rough idea about the likely shape of the underlying probability distribution.

[15]For a detailed discussion of various normality tests, see G. Barrie Wetherhill, *Regression Analysis with Applications,* Chapman and Hall, London, 1986, Chap. 8.

Normal Probability Plot

Another comparatively simple graphical device to study the PDF of a random variable is the **normal probability plot (NPP)** which makes use of *normal probability paper,* a specially ruled graph paper. On the horizontal axis, (X-axis) we plot values of the variable of interest (say, OLS residuals e_i), and on the vertical axis (Y-axis), we show the expected values of this variable if its distribution were normal. Therefore, if the variable is in fact from the normal population, the NPP will approximate a straight line. MINITAB has the capability to plot the NPP of any random variable. MINITAB also produces the **Anderson-Darling normality test** known as the A^2 **statistic.** The underlying null hypothesis is that a variable is normally distributed. This hypothesis can be sustained if the computed A^2 is not statistically significant.

Jarque-Bera Test

A test of normality that has now become very popular and is included in several statistical packages is the **Jarque-Bera (JB) test.**[16] This is an *asymptotic,* or large sample, *test* and is based on OLS residuals. This test first computes the coefficients of *skewness, S* (a measure of asymmetry of a PDF), and *kurtosis, K* (a measure of how tall or flat a PDF is in relation to the normal distribution), of a random variable (e.g., OLS residuals) (see Appendix B). For a normally distributed variable, skewness is zero and kurtosis is 3 (see Figure B-4 in Appendix B).

Jarque and Bera have developed the following test statistic:

$$\text{JB} = \frac{n}{6}\left[S^2 + \frac{(K-3)^2}{4}\right] \tag{3.47}$$

where n is the sample size, S represents skewness, and K represents kurtosis. They have shown that under the normality assumption the JB statistic given in Equation (3.47) *follows the chi-square distribution with 2 d.f. asymptotically* (i.e., in large samples). Symbolically,

$$\text{JB}_{\text{asy}} \sim \chi^2_{(2)} \tag{3.48}$$

where asy means asymptotically.

As you can see from Eq. (3.47), if a variable is normally distributed, S is zero and $(K-3)$ is also zero, and therefore the value of the JB statistic is zero ipso facto. But if a variable is not normally distributed, the JB statistic will assume increasingly larger values. What constitutes a large or small value of the JB statistic can be learned easily from the chi-square table (Appendix E, Table E-4). If the computed chi-square value from Eq. (3.47) exceeds the critical chi-square value for 2 d.f. at the chosen level of significance, we reject the null hypothesis of normal distribution; but if it does not exceed the critical chi-square value, we do

[16]See C. M. Jarque and A. K. Bera, "A Test for Normality of Observations and Regression Residuals," *International Statistical Review,* vol. 55, 1987, pp. 163–172.

not reject the null hypothesis. Of course, if we have the *p* value of the computed chi-square value, we will know the exact probability of obtaining that value.

We will illustrate these normality tests with the following example.

3.10 A CONCLUDING EXAMPLE: RELATIONSHIP BETWEEN WAGES AND PRODUCTIVITY IN THE U.S. BUSINESS SECTOR, 1959–2006

According to the marginal productivity theory of microeconomics, we would expect a positive relationship between wages and worker productivity. To see if this so, in Table 3-3 (on the textbook's Web site) we provide data on labor productivity, as measured by the index of output per hour of all persons, and wages, as measured by the index of real compensation per hour, for the business sector of the U.S. economy for the period 1959 to 2006. The base year of the index is 1992 and hourly real compensation is hourly compensation divided by the consumer price index (CPI).

Let *Compensation* (Y) = index of real compensation and *Productivity* (X) = index of output per hour of all persons. Plotting these data, we obtain the scatter diagram shown in Figure 3-10.

This figure shows a very close linear relationship between labor productivity and real wages. Therefore, we can use a (bivariate) linear regression to

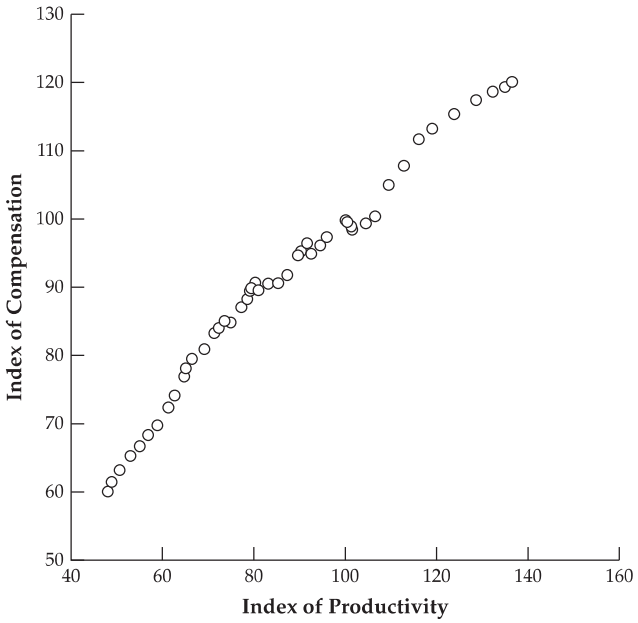

FIGURE 3-10 Relationship between compensation and productivity in the U.S. business sector, 1959–2006

model the data given in Table 3-3. Using EViews, we obtain the following results:

Dependent Variable: Compensation
Method: Least Squares
Sample: 1959 2006
Included observations: 48

	Coefficient	Std. Error	*t*-Statistic	Prob.
C	33.63603	1.400085	24.02428	0.0000
Productivity	0.661444	0.015640	42.29178	0.0000

R-squared	0.974926
Adjusted R-squared	0.974381
S.E. of regression	2.571761
Sum squared resid	304.2420
Durbin-Watson stat	0.146315

Let us interpret the results. The slope coefficient of about 0.66 suggests that if the index of productivity goes up by a unit, the index of real wages will go up, on average, by 0.66 units. This coefficient is highly significant, for the t value of about 42.3 (obtained under the assumption that the true population coefficient is zero) is highly significant for the p value is almost zero. The intercept coefficient, C, is also highly significant, for the p value of obtaining a t value for this coefficient of as much as about 24 is practically zero.

The R^2 value of about 0.97 means that the index of productivity explains about 97 percent of the variation in the index of real compensation. This is a very high value, since an R^2 can at most be 1. For now neglect some of the information given in the preceding table (e.g., the Durbin-Watson statistic), for we will explain it at appropriate places.

Figure 3-11 gives the actual and estimated values of the index of real compensation, the dependent variable in our model, as well the differences between the two, which are nothing but the residuals e_i. These residuals are also plotted in this figure.

Figure 3-12 plots the histogram of the residuals shown in Figure 3-11 and also shows the JB statistics. The histogram and the JB statistic show that there is no reason to reject the hypothesis that the true error terms in the wages-productivity regression are normally distributed.

Figure 3-13 shows the normal probability plot of the residuals obtained from the compensation-productivity regression; this figure was obtained from MINITAB. As is clear from this figure, the estimated residuals lie approximately on a straight line, suggesting that the error terms (i.e., u_i) in this regression may be normally distributed. The computed AD statistic of 0.813 has a p value of about 0.03 or 3 percent. If we fix the critical p value, say, at the 5 percent level, the observed AD statistic is statistically significant, suggesting that the error terms are not normally distributed. This is in contrast to the conclusion reached on the basis of the JB

Actual Y_i	Fitted $\hat{Y}_i$	Residual e_i	Residual plot (−) (0) (+)
59.8710	65.4025	−5.53155	
61.3180	65.9575	−4.63950	
63.0540	67.0833	−4.02928	
65.1920	68.6145	−3.42252	
66.6330	69.9824	−3.34939	
68.2570	71.2113	−2.95435	
69.6760	72.5402	−2.86419	
72.3000	74.1191	−1.81906	
74.1210	75.0041	−0.88307	
76.8950	76.4163	0.47875	
78.0080	76.6253	1.38273	
79.4520	77.4799	1.97214	
80.8860	79.2856	1.60040	
83.3280	80.7593	2.56870	
85.0620	82.1926	2.86936	
83.9880	81.4300	2.55800	
84.8430	83.1068	1.73624	
87.1480	84.6631	2.48486	
88.3350	85.5296	2.80537	
89.7360	86.1018	3.63422	
89.8630	86.0919	3.77114	
89.5920	85.9900	3.60200	
89.6450	87.0662	2.57884	
90.6370	86.6495	3.98755	
90.5910	88.5366	2.05445	
90.7120	90.0003	0.71167	
91.9100	91.2683	0.64168	
94.8690	92.9497	1.91929	
95.2070	93.2540	1.95303	
96.5270	94.1621	2.36486	
95.0050	94.7588	0.24624	
96.2190	96.0664	0.15257	
97.4650	97.0705	0.39449	
100.0000	99.7804	0.21956	
99.7120	100.0360	−0.32376	
99.0240	100.6730	−1.64873	
98.6900	100.7690	−2.07930	
99.4780	102.7520	−3.27365	
100.5120	104.0650	−3.55328	
105.1730	106.0470	−0.87396	
108.0440	108.2650	−0.22145	
111.9920	110.4410	1.55106	
113.5360	112.4020	1.13388	
115.6940	115.6210	0.07329	
117.7090	118.7670	−1.05820	
118.9490	121.2050	−2.25562	
119.6920	122.9450	−3.25288	
120.4470	123.8600	−3.41265	

FIGURE 3-11 Actual Y, estimated Y, and residuals (regression of compensation on productivity)

Note: Y = Actual index of compensation

$\hat{Y}$ = Estimated index of compensation

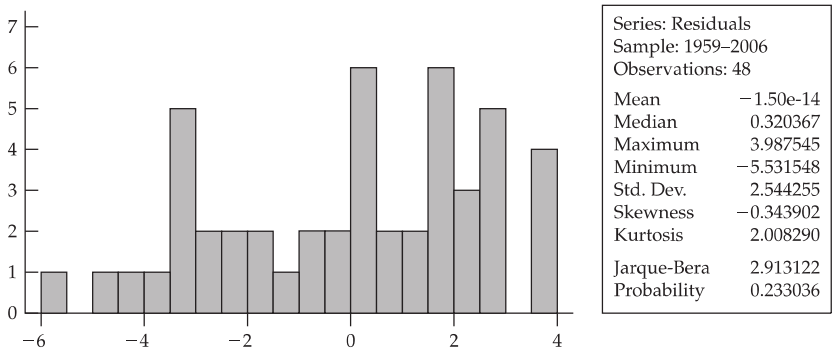

FIGURE 3-12 Histogram of residuals from the compensation-productivity regression

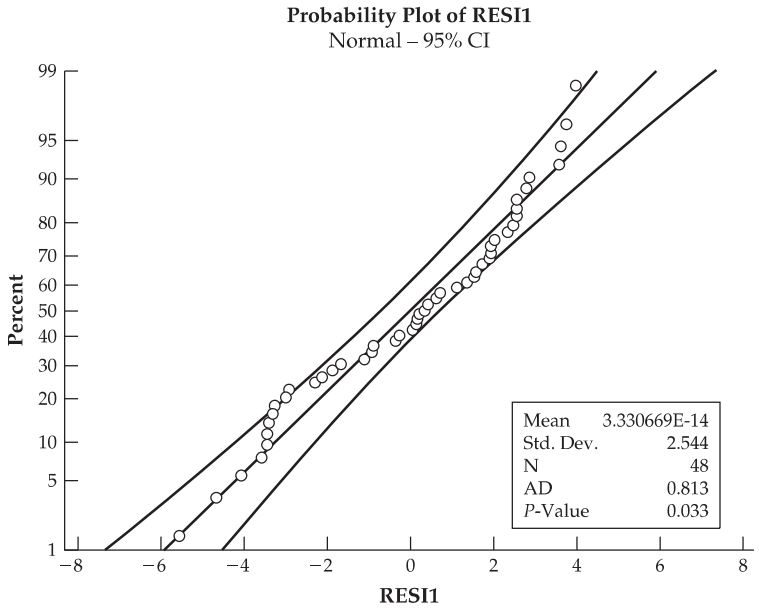

FIGURE 3-13 Normal probability plot of residuals obtained from the compensation-productivity regression

statistic. The problem here is that our sample of 10 observations is too small for using the JB and AD statistics, which are designed for large samples.

3.11 A WORD ABOUT FORECASTING

We noted in Chapter 2 that one of the purposes of regression analysis is to predict the mean value of the dependent variable, given the values of the explanatory variable(s). To be more specific, let us return to our math S.A.T. score example. Regression (3.46) presented the results of the math section of the S.A.T. based on the score data of Table 2-2. Suppose we want to find out the

average math S.A.T. score by a person with a given level of annual family income. What is the expected math S.A.T. score at this level of annual family income?

To fix these ideas, assume that X (income) takes the value X_0, where X_0 is some *specified numerical value* of X, say $X_0 = \$78{,}000$. Now suppose we want to estimate $E(Y \mid X_0 = 78000)$, that is, the true mean math S.A.T. score corresponding to a family income of \$78,000. Let

$$\hat{Y}_0 = \text{the estimator of } E(Y \mid X_0) \tag{3.49}$$

How do we obtain this estimate? Under the assumptions of the classical linear regression model (CLRM), it can be shown that Equation (3.49) can be obtained by simply putting the given X_0 value in Eq. (3.46), which gives:

$$\begin{aligned} \hat{Y}_{X=78000} &= 432.4138 + 0.0013(78000) \\ &= 533.8138 \end{aligned} \tag{3.50}$$

That is, the forecasted mean math S.A.T. score for a person with an annual family income of \$78,000 is about 534 points.

Although econometric theory shows that under CLRM $\hat{Y}_{\hat{X}=78000}$, or, more generally, $\hat{Y}_0$ is an unbiased estimator of the true mean value (i.e., a point on the population regression line), it is not likely to be equal to the latter in any given sample. (Why?) The difference between them is called the **forecasting,** or **prediction, error.** To assess this error, we need to find out the sampling distribution of $\hat{Y}_0$.[17] Given the assumptions of the CLRM, it can be shown that $\hat{Y}_0$ is *normally distributed* with the following mean and variance:

$$\begin{aligned} \text{Mean} &= E(Y \mid X_0) = B_1 + B_2 X_0 \\ \text{var} &= \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x_i^2} \right] \end{aligned} \tag{3.51}$$

where $\overline{X}$ = the sample mean of X values in the historical regression (3.46)
$\sum x_i^2$ = their sum of squared deviations from $\overline{X}$
σ^2 = the variance of u_i
n = sample size

The positive square root of Equation (3.51) gives the standard error of $\hat{Y}_0$, $\text{se}(\hat{Y}_0)$.

Since in practice σ^2 is not known, if we replace it by its unbiased estimator $\hat{\sigma}^2$, $\hat{Y}_0$ follows the t distribution with $(n - 2)$ d.f. (Why?) Therefore, we can use the t distribution to establish a $100\ (1 - \alpha)\%$ confidence interval for the true (i.e., population) mean value of Y corresponding to X_0 in the usual manner as follows:

$$P[b_1 + b_2 X_0 - t_{a/2}\,\text{se}(\hat{Y}_0) \leq B_1 + B_2 X_0 \leq b_1 + b_2 X_0 + t_{a/2}\,\text{se}(\hat{Y}_0)] = (1 - \alpha) \tag{3.52}$$

[17] Note that $\hat{Y}_0$ is an estimator and therefore will have a sampling distribution.

Let us continue with our math S.A.T. score example. First, we compute the variance of $\hat{Y}_{X=78000}$ from Equation (3.51).

$$\begin{aligned} \text{var}\left(\hat{Y}_{X=78000}\right) &= 975.1347\left[\frac{1}{10} + \frac{(78{,}000 - 56{,}000)^2}{16{,}240{,}000{,}000}\right] \\ &= 126.5754 \end{aligned} \tag{3.53}$$

Therefore,

$$\begin{aligned} \text{se}\left(\hat{Y}_{X=78000}\right) &= \sqrt{126.5754} \\ &= 11.2506 \end{aligned} \tag{3.54}$$

Note: In this example, $\overline{X} = 56000$, $\sum x_i^2 = 16{,}240{,}000{,}000$, and $\hat{\sigma}^2 = 975.1347$ (see Table 2-4).

The preceding result suggests that given the estimated annual family income = $78,000, the mean predicted math S.A.T. score, as shown in Equation (3.50), is 533.8138 points and the standard error of this predicted value is 11.2506 (points).

Now if we want to establish, say, a 95% confidence interval for the population mean math S.A.T. score corresponding to an annual family income of $78,000, we obtain it from expression (3.52) as

$$533.8138 - 2.306(11.2506) \leq E(Y \mid X = 78000) \leq 533.8138 + 2.306\,(11.2506)$$

That is,

$$507.8699 \leq E(Y \mid X = 78000) \leq 559.7577 \tag{3.55}$$

Note: For 8 d.f., the 5 percent two-tailed t value is 2.306.

Given the annual family income of $78,000, Equation (3.55) states that although the single best, or point, estimate of the mean math S.A.T. score is 533.8138, it is expected to lie in the interval 507.8699 to 559.7577 points, which is between about 508 and 560, with 95% confidence. Therefore, with 95% confidence, the forecast error will be between −25.9439 points (507.8699 − 533.8138) and 25.9439 points (559.7577 – 533.8138).

If we obtain a 95% confidence interval like Eq. (3.55) for each value of X shown in Table 2-2, we obtain what is known as a **confidence interval** or **confidence band** for the true mean math S.A.T. score for each level of annual family income, or for the entire population regression line (PRL). This can be seen clearly from Figure 3-14, obtained from EViews.

Notice some interesting aspects of Figure 3-14. The width of the confidence band is smallest when $X_0 = \overline{X}$, which should be apparent from the variance formula given in Eq. (3.51). However, the width widens sharply (i.e.,

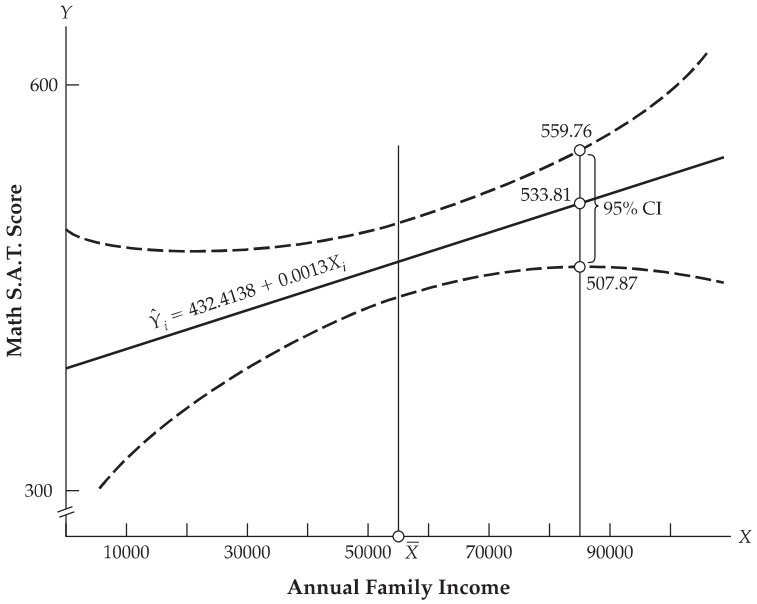

FIGURE 3-14 95% confidence band for the true math S.A.T. score function

the prediction error increases) as X_0 moves away from $\overline{X}$. This suggests that the predictive ability of the historical regression, such as regression (3.46), falls markedly as X_0 (the X value for which the forecast is made) departs progressively from $\overline{X}$. The message here is clear: *We should exercise great caution in "extrapolating" the historical regression line to predict the mean value of Y associated with any X that is far removed from the sample mean of X. In more practical terms, we should not use the math S.A.T. score regression (3.46) to predict the average math score for income well beyond the sample range on which the historical regression line is based.*

3.12 SUMMARY

In Chapter 2 we showed how to estimate the parameters of the two-variable linear regression model. In this chapter we showed how the estimated model can be used for the purpose of drawing inferences about the true population regression model. Although the two-variable model is the simplest possible linear regression model, the ideas introduced in these two chapters are the foundation of the more involved multiple regression models that we will discuss in ensuing chapters. As we will see, in many ways the multiple regression model is a straightforward extension of the two-variable model.

KEY TERMS AND CONCEPTS

The key terms and concepts introduced in this chapter are

Classical linear regression model (CLRM)
Homoscedasticity or equal variance
Heteroscedasticity or unequal variance
Autocorrelation and no autocorrelation
Variances of OLS estimators
Standard errors of OLS estimators
Residual sum of squares (RSS)
Standard error of the regression (SER)
Sampling, or probability, distributions of OLS estimators
Gauss-Markov theorem
BLUE property
Central limit theorem (CLT)
"Zero" null hypothesis; straw man hypothesis
t test of significance
  a) two-tailed t test
  b) one-tailed t test
Coefficient of determination, r^2
Total sum of squares (TSS)
Explained sum of squares (ESS)
Coefficient of alienation
Coefficient of correlation, r
Normal probability plot (NPP)
Anderson-Darling normality test (A^2 statistic)
Jarque-Bera (JB) test of normality
Forecasting, or prediction, error
Confidence interval; confidence band

QUESTIONS

3.1. Explain the meaning of
  a. Least squares.
  b. OLS estimators.
  c. The variance of an estimator.
  d. Standard error of an estimator.
  e. Homoscedasticity.
  f. Heteroscedasticity.
  g. Autocorrelation.
  h. Total sum of squares (TSS).
  i. Explained sum of squares (ESS).
  j. Residual sum of squares (RSS).
  k. r^2.
  l. Standard error of estimate.
  m. BLUE.
  n. Test of significance.
  o. t test.
  p. One-tailed test.
  q. Two-tailed test.
  r. Statistically significant.

3.2. State with brief reasons whether the following statements are true, false, or uncertain.
  a. OLS is an estimating procedure that minimizes the sum of the errors squared, $\sum u_i^2$.
  b. The assumptions made by the classical linear regression model (CLRM) are not necessary to compute OLS estimators.

c. The theoretical justification for OLS is provided by the Gauss-Markov theorem.
d. In the two-variable PRF, b_2 is likely to be a more accurate estimate of B_2 if the disturbances u_i follow the normal distribution.
e. The OLS estimators b_1 and b_2 each follow the normal distribution only if u_i follows the normal distribution.
f. r^2 is the ratio of TSS/ESS.
g. For a given alpha and d.f., if the computed $|t|$ exceeds the critical t value, we should accept the null hypothesis.
h. The coefficient of correlation, r, has the same sign as the slope coefficient b_2.
i. The p value and the level of significance, α, mean the same thing.

3.3. Fill in the appropriate gaps in the following statements:
a. If $B_2 = 0$, $b_2/\text{se}(b_2) = \ldots$
b. If $B_2 = 0$, $t = b_2/\ldots$
c. r^2 lies between . . . and . . .
d. r lies between . . . and . . .
e. TSS = RSS + . . .
f. d.f. (of TSS) = d.f. (of . . .) + d.f. (of RSS)
g. $\hat{\sigma}$ is called . . .
h. $\sum y_i^2 = \sum(Y_i - \ldots)^2$
i. $\sum y_i^2 = b_2(\ldots)$

3.4. Consider the following regression:

$$\begin{aligned} \hat{Y}_i &= -66.1058 + 0.0650X_i & r^2 &= 0.9460 \\ \text{se} &= (10.7509) \quad (\quad) & n &= 20 \\ t &= (\quad) \quad (18.73) \end{aligned}$$

Fill in the missing numbers. Would you reject the hypothesis that true B_2 is zero at $\alpha = 5\%$? Tell whether you are using a one-tailed or two-tailed test and why.

3.5. Show that all the following formulas to compute r^2 are equivalent:

$$\begin{aligned} r^2 &= 1 - \frac{\sum e_i^2}{\sum y_i^2} \\ &= \frac{\sum \hat{y}_i^2}{\sum y_i^2} \\ &= \frac{b_2^2 \sum x_i^2}{\sum y_i^2} \\ &= \frac{\left(\sum y_i \hat{y}_i\right)^2}{\left(\sum y_i^2\right)\left(\sum \hat{y}_i^2\right)} \end{aligned}$$

3.6. Show that $\sum e_i = n\overline{Y} - nb_1 - nb_2\overline{X} = 0$

PROBLEMS

3.7. Based on the data for the years 1962 to 1977 for the United States, Dale Bails and Larry Peppers[18] obtained the following demand function for automobiles:

$$\hat{Y}_t = 5807 + 3.24X_t \qquad r^2 = 0.22$$
$$\text{se} = \qquad (1.634)$$

where Y = retail sales of passenger cars (thousands) and X = the real disposable income (billions of 1972 dollars).
Note: The se for b_1 is not given.

a. Establish a 95% confidence interval for B_2.
b. Test the hypothesis that this interval includes $B_2 = 0$. If not, would you accept this null hypothesis?
c. Compute the t value under H_0:$B_2 = 0$. Is it statistically significant at the 5 percent level? Which t test do you use, one-tailed or two-tailed, and why?

3.8. The *characteristic line* of modern investment analysis involves running the following regression:

$$r_1 = B_1 + B_2 r_{mt} + u_t$$

where r = the rate of return on a stock or security
r_m = the rate of return on the market portfolio represented by a broad market index such as S&P 500, and
t = time

In investment analysis, B_2 is known as the *beta coefficient* of the security and is used as a measure of market risk, that is, how developments in the market affect the fortunes of a given company.

Based on 240 monthly rates of return for the period 1956 to 1976, Fogler and Ganapathy obtained the following results for IBM stock. The market index used by the authors is the market portfolio index developed at the University of Chicago:[19]

$$r_t = 0.7264 + 1.0598 r_{mt}$$
$$\text{se} = (0.3001)\ (0.0728) \qquad r^2 = 0.4710$$

a. Interpret the estimated intercept and slope.
b. How would you interpret r^2?
c. A security whose beta coefficient is greater than 1 is called a volatile or aggressive security. Set up the appropriate null and alternative hypotheses and test them using the t test. *Note*: Use $\alpha = 5\%$.

3.9. You are given the following data based on 10 pairs of observations on Y and X.

$$\sum y_i = 1110 \quad \sum X_i = 1680 \quad \sum X_i Y_i = 204{,}200$$
$$\sum X_i^2 = 315{,}400 \quad \sum Y_i^2 = 133{,}300$$

[18]See Dale G. Bails and Larry C. Peppers, *Business Fluctuations: Forecasting Techniques and Applications,* Prentice-Hall, Englewood Cliffs, N.J., 1982, p. 147.
[19]H. Russell Fogler and Sundaram Ganapathy, *Financial Econometrics,* Prentice-Hall, Englewood-Cliffs, N.J., 1982, p. 13.

Assuming all the assumptions of CLRM are fulfilled, obtain

a. b_1 and b_2.
b. standard errors of these estimators.
c. r^2.
d. Establish 95% confidence intervals for B_1 and B_2.
e. On the basis of the confidence intervals established in (*d*), can you accept the hypothesis that $B_2 = 0$?

3.10. Based on data for the United States for the period 1965 to 2006 (found in Table 3-4 on the textbook's Web site), the following regression results were obtained:

$$\text{GNP}_t = -995.5183 + 8.7503M_{1t} \quad r^2 = 0.9488$$
$$\text{se} = (\quad) \qquad (0.3214)$$
$$t = (-3.8258) \qquad (\quad)$$

where GNP is the gross national product ($, in billions) and M_1 is the money supply ($, in billions).

Note: M_1 includes currency, demand deposits, traveler's checks, and other checkable deposits.

a. Fill in the blank parentheses.
b. The monetarists maintain that money supply has a significant positive impact on GNP. How would you test this hypothesis?
c. What is the meaning of the negative intercept?
d. Suppose M_1 for 2007 is $750 billion. What is the mean forecast value of GNP for that year?

3.11. *Political business cycle:* Do economic events affect presidential elections? To test this so-called political business cycle theory, Gary Smith[20] obtained the following regression results based on the U.S. presidential elections for the four yearly periods from 1928 to 1980 (i.e., the data are for years 1928, 1932, etc.):

$$\hat{Y}_t = 53.10 - 1.70X_t$$
$$t = (34.10)\ (-2.67) \qquad r^2 = 0.37$$

where Y is the percentage of the vote received by the incumbent and X is the unemployment rate change—unemployment rate in an election year minus the unemployment rate in the preceding year.

a. A priori, what is the expected sign of X?
b. Do the results support the political business cycle theory? Support your contention with appropriate calculations.
c. Do the results of the 1984 and 1988 presidential elections support the preceding theory?
d. How would you compute the standard errors of b_1 and b_2?

3.12. To study the relationship between capacity utilization in manufacturing and inflation in the United States, we obtained the data shown in Table 3-5 (found on the textbook's Web site). In this table, Y = inflation rate as measured by the

[20]Gary Smith, *Statistical Reasoning,* Allyn & Bacon, Boston, Mass., 1985, p. 488. Change in notation was made to conform with our format. The original data were obtained by Ray C. Fair, "The Effect of Economic Events on Votes for President," *The Review of Economics and Statistics,* May 1978, pp. 159–173.

percentage change in GDP implicit price deflator and X = capacity utilization rate in manufacturing as measured by output as a percent of capacity for the years 1960–2007.

a. A priori, what would you expect to be the relationship between inflation rate and capacity utilization rate? What is the economic rationale behind your expectation?
b. Regress Y on X and present your result in the format of Eq. (3.46).
c. Is the estimated slope coefficient statistically significant?
d. Is it statistically different from unity?
e. The natural rate of capacity utilization is defined as the rate at which Y is zero. What is this rate for the period under study?

3.13. *Reverse regression*[21]: Continue with Problem 3.12, but suppose we now regress X on Y.

a. Present the result of this regression and comment.
b. If you multiply the slope coefficients in the two regressions, what do you obtain? Is this result surprising to you?
c. The regression in Problem 3.12 may be called the *direct regression.* When would a reverse regression be appropriate?
d. Suppose the r^2 value between X and Y is 1. Does it then make any difference if we regress Y on X or X on Y?

3.14. Table 3-6 gives data on X (net profits after tax in U.S. manufacturing industries [$, in millions]) and Y (cash dividend paid quarterly in manufacturing industries [$, in millions]) for years 1974 to 1986.

a. What relationship, if any, do you expect between cash dividend and after-tax profits?
b. Plot the scattergram between Y and X.
c. Does the scattergram support your expectations in part (*a*)?
d. If so, do an OLS regression of Y on X and obtain the usual statistics.
e. Establish a 99% confidence interval for the true slope and test the hypothesis that the true slope coefficient is zero; that is, there is no relationship between dividend and the after-tax profit.

TABLE 3-6 CASH DIVIDEND (Y) AND AFTER-TAX PROFITS (X) IN U.S. MANUFACTURING INDUSTRIES, 1974–1986

Year	Y	X	Year	Y	X
	($, in millions)			($, in millions)	
1974	19,467	58,747	1981	40,317	101,302
1975	19,968	49,135	1982	41,259	71,028
1976	22,763	64,519	1983	41,624	85,834
1977	26,585	70,366	1984	45,102	107,648
1978	28,932	81,148	1985	45,517	87,648
1979	32,491	98,698	1986	46,044	83,121
1980	36,495	92,579			

Source: Business Statistics, 1986, U.S. Department of Commerce, Bureau of Economic Analysis, December 1987, p. 72.

[21]On this see G. S. Maddala, *Introduction to Econometrics,* 3rd ed., Wiley, New York, 2001, pp. 71–75.

3.15. Refer to the S.A.T. data given in Table 2-15 on the textbook's Web site. Suppose you want to predict the male math scores on the basis of the female math scores by running the following regression:

$$Y_t = B_1 + B_2 X_t + u_t$$

where Y and X denote the male and female math scores, respectively.

a. Estimate the preceding regression, obtaining the usual summary statistics.
b. Test the hypothesis that there is no relationship between Y and X whatsoever.
c. Suppose the female math score in 2008 is expected to be 490. What is the predicted (average) male math score?
d. Establish a 95% confidence interval for the predicted value in part (*c*).

3.16. Repeat the exercise in Problem 3.15 but let Y and X denote the male and the female critical reading scores, respectively. Assume a female critical reading score for 2008 of 505.

3.17. Consider the following regression results:[22]

$$\hat{Y}_t = -0.17 + 5.26X_t \qquad \overline{R}^2 = 0.10, \text{ Durbin-Watson} = 2.01$$
$$t = (-1.73)(2.71)$$

where Y = the real return on the stock price index from January of the current year to January of the following year
X = the total dividends in the preceding year divided by the stock price index for July of the preceding year
t = time

Note: On Durbin-Watson statistic, see Chapter 10.
The time period covered by the study was 1926 to 1982.
Note: $\overline{R}^2$ stands for the adjusted coefficient of determination. The Durbin-Watson value is a measure of autocorrelation. Both measures are explained in subsequent chapters.

a. How would you interpret the preceding regression?
b. If the previous results are acceptable to you, does that mean the best investment strategy is to invest in the stock market when the dividend/price ratio is high?
c. If you want to know the answer to part (*b*), read Shiller's analysis.

3.18. Refer to Example 2.1 on years of schooling and average hourly earnings. The data for this example are given in Table 2-5 and the regression results are presented in Eq. (2.21). For this regression

a. Obtain the standard errors of the intercept and slope coefficients and r^2.
b. Test the hypothesis that schooling has no effect on average hourly earnings. Which test did you use and why?
c. If you reject the null hypothesis in (*b*), would you also reject the hypothesis that the slope coefficient in Eq. (2.21) is not different from 1? Show the necessary calculations.

[22]See Robert J. Shiller, *Market Volatility*, MIT Press, Cambridge, Mass., 1989, pp. 32–36.

3.19. Example 2.2 discusses Okun's law, as shown in Eq. (2.22). This equation can also be written as $X_t = B_1 + B_2 Y_t$, where X = percent growth in real output, as measured by GDP and Y = change in the unemployment rate, measured in percentage points. Using the data given in Table 2-13 on the textbook's Web site,

- **a.** Estimate the preceding regression, obtaining the usual results as per Eq. (3.46).
- **b.** Is the change in the unemployment rate a significant determinant of percent growth in real GDP? How do you know?
- **c.** How would you interpret the intercept coefficient in this regression? Does it have any economic meaning?

3.20. For Example 2.3, relating stock prices to interest rates, are the regression results given in Eq. (2.24) statistically significant? Show the necessary calculations.

3.21. Refer to Example 2.5 about antique clocks and their prices. Based on Table 2-14, we obtained the regression results shown in Eqs. (2.27) and (2.28). For each regression obtain the standard errors, the t ratios, and the r^2 values. Test for the statistical significance of the estimated coefficients in the two regressions.

3.22. Refer to Problem 3.22. Using OLS regressions, answer questions (*a*), (*b*), and (*c*).

3.23. Table 3-7 (found on the textbook's Web site) gives data on U.S. expenditure on imported goods (Y) and personal disposable income (X) for the period 1959 to 2006.

Based on the data given in this table, estimate an import expenditure function, obtaining the usual regression statistics, and test the hypothesis that expenditure on imports is unrelated to personal disposable income.

3.24. Show that the OLS estimators, b_1 and b_2, are linear estimators. Also show that these estimators are linear functions of the error term u_i (*Hint:* Note that $b_2 = \sum x_i y_i / \sum x_i^2 = \sum w_i y_i$, where $w_i = x_i / \sum x_i^2$ and note that the X's are nonstochastic).

3.25. Prove Eq. (3.35). (*Hint:* Square Eq. [3.33] and use some of the properties of OLS).