3

Model Performance and Evaluation

In the previous chapter, we learned how to use the method of least-squares to find a line that best fits a scatter of points. It can be shown that, given the right circumstances, there is no other method that is better at estimating such a line than least-squares. That is, if the CLRM assumptions are met, then OLS is BLUE. At this point, two other issues arise. First, suppose we have used least-squares to find a line that best fits the data. We can then ask *how well* the line fits the data. Finding the best fit is one thing, but how well our line fits the data is quite another. This is the issue of “goodness of fit,” which we explore in the next section.

Another question we can ask is not how well the regression model as a whole performs, but how do the separate pieces of the model perform? That is, in the previous chapter, we learned how to calculate the least-squares values of a and b to determine the sample regression line. But because these values were derived from a sample, they may not be representative of the *population's* α and β. Thus, we can ask the question, how confident are we that our sample results are a good reflection of the population's behavior? We discuss this issue later in the chapter.

Goodness of Fit: The R^2

In Chapter 2, we estimated four sample regression lines, one for each of our examples. We then plotted these lines along with the actual data for the dependent variable, *Y*, and the independent variable, *X*. We have already accepted the fact that in each case, the dots representing the sample data will not fall exactly on the sample regression line, reflecting the fact that our model does not take into account all factors that affect our dependent variable. In general, however, we hope that the vertical distances from the dots to the regression line are small because if this is true, then our estimated line would be a good predictor of the behavior of *Y*.

Reviewing Figures 2.3, 2.4, 2.5, and 2.6a from the previous chapter, we can see some qualitative differences. In Figure 2.3, which is for our baseball salary example, we see that the plotted data points are somewhat broadly scattered around the sample regression line. This indicates that the model explains some of the behavior of *Y*, but much is left unexplained. Comparing this to Figure 2.4, which is for the presidential voting model, we see that the plotted points in this case are more closely cropped around the regression line. Thus, for this example, the regression line seems to tell us a lot about the behavior of the dependent variable, *Y*. Finally, in Figures 2.5 and 2.6a, we see that the plotted data for state abortion rates and crime in California are spread very broadly around their respective regression lines. In the case of abortion, we can conclude that although religion may be an important factor in determining abortion rates, there are many other important factors we need to consider. The same can be said for the case of British crime rates: Unemployment rates may be one of the determining factors, but there are obviously other causal factors not present in the model.

Even though the least-squares regression method produces the best possible line to fit our data,[1] this means only that we have done the best we can, and the overall performance of the model is not guaranteed to be good. In order to judge how well the model fits the data, we can employ a measure called the $\boldsymbol{R}^2$ (read as the ***R* squared**).[2] The technical derivation of this measure can be a little hard to follow, but the intuition of it is not too difficult to understand.

Our basic approach, as set out in Chapter 1, is to understand the behavior of *Y*, the dependent variable, by observing the behavior of *X*, the independent variable. As the value of *X* differs from one observation to another, we expect that this difference in *X* will explain, at least in part, the differences in *Y* from one observation to another. This approach is summarized by Equation 1.5.

[1] Again, this will be true only if the CLRM assumptions are met. As we see in later chapters, our simple models may not satisfy all of the necessary conditions that are needed before we can say that we have found the best possible line to fit our data. In order to continue with our discussion, however, we will assume that all conditions have been satisfied and our least-squares estimates are the best possible.

[2] The R^2 is also referred to as the "coefficient of determination."

We hope that by observing the behavior of X, we learn a great deal about the behavior of Y. The behavior of Y that is not explained by X will be captured by the error term in Equation 1.5. *The R^2 simply is a measure that tells us what proportion of the behavior of Y is explained by X.*

We can easily consider the bounds for the R^2. If, by observing the behavior of X, we know exactly how Y behaves, then the R^2 would be equal to 1 (100%). This outcome is very unlikely.[3] On the other hand, observing X may tell us nothing about Y, in which case the R^2 would be equal to 0 (0%). Thus, the R^2 is bounded between 0 and 1. Values close to 1 mean that by observing the behavior of X, we can explain nearly all of the behavior of Y. This would indicate that our estimated sample regression function is performing well overall. Values close to 0 would have the opposite implication.

The above discussion gives us an intuitive understanding of the R^2. In order to understand the technical derivation of this measure, we need to define what is meant by the "behavior" of Y. The dependent variable Y will have a mean (average), and of course, some values of Y will fall above this mean and some below. As a graphic example, we can consider Figure 3.1.

As we can see in Figure 3.1, a particular point is plotted for observation Y_i for a given value of X_i. Of course, there would normally be many other observations plotted on the graph, but they are not shown in this case so that we may focus on this single observation. The mean value for the dependent variable, $\bar{Y}$, is a horizontal line. The sample regression function (SRF) is also plotted. Notice that the plotted point Y_i lies above $\bar{Y}$ (i.e., it is above the horizontal line). Thus, the value of Y for this particular observation is above the sample's average value for Y. The amount by which Y_i exceeds $\bar{Y}$ is shown as the distance from the horizontal line to the point Y_i and is denoted as $(Y_i - \bar{Y})$. This deviation of Y_i from its mean can be broken up into two pieces: the amount of the deviation that is predicted by our model, and the amount that our model does not predict. That part of the deviation that our model predicts is shown as the vertical distance from the mean of Y $(\bar{Y})$ to the value the model would predict for Y $(\hat{Y}_i)$ for the given X_i. This distance is denoted $(\hat{Y}_\text{i} - \bar{Y})$. The second piece is the part of the total deviation that was not predicted by our model, shown as the vertical distance from the point on the sample regression function, $\hat{Y}_i$, to the observation Y_i, and this distance is denoted as $(Y_i - \hat{Y}_i)$, which is simply our error term, e_i. Thus, suppose this observation is from our baseball example. Then the salary for player *i*, Y_i, is above the average salary for all players in our sample. Part of the amount by which player *i*'s salary exceeds the average is explained by our model. That is, it can be explained by the number of years player *i* has

[3] If the R^2 turns out to be equal to 1, the most likely reason is that the model that was estimated was, in fact, an identity. (See footnote 4 in Chapter 1.)

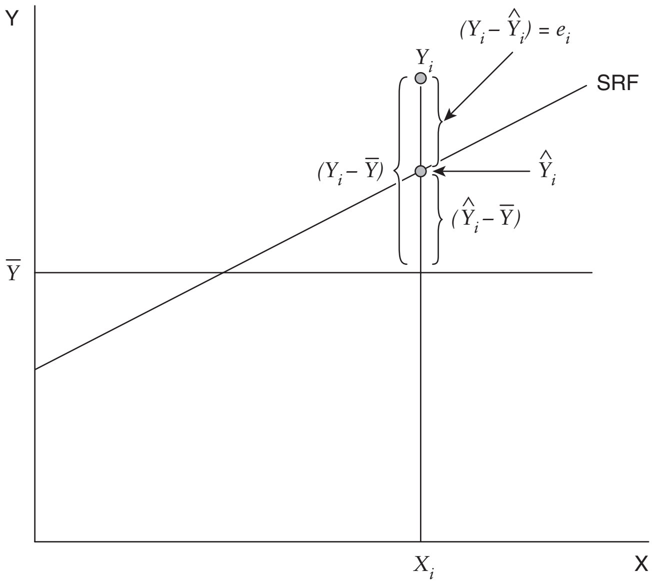

Figure 3.1

played in MLB. The rest is not explained by our model and is captured by the residual, e_i. It should be obvious that the greater the proportion of the deviation of the observation from its mean value that is explained by our model, the better our model is performing because this would mean that the proportion accounted for by the error term is smaller. On the other hand, if our model does a poor job of explaining the deviation of Y_i from its mean, then our error term will be larger. This kind of breakdown of the total deviation of Y_i from its mean into the explained portion and the unexplained portion can be done for all observations in our sample.

Now, as noted above, the R^2 is a measure that tells us what proportion of the behavior of *Y* is explained by *X*. We can now use the method described above to give a more precise meaning to the "behavior of *Y*." We can define this behavior of *Y* as the *variation* in *Y*, and it is calculated according to the following equation:

$$\text{TSS} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2. \tag{3.1}$$

That is, the variation in *Y* is the sum of the squared deviations of *Y* around its mean. We will call this sum the "Total Sum of Squares," or TSS for short. Essentially, it is a measure that tells us by how much the values of *Y* "bounce" around its mean. The deviations are squared so as to prevent cancellation of positive values with negative ones (recall that this was also done when we derived the OLS method in Chapter 2).

Part of this behavior of Y (i.e., the TSS) is explained or "predicted" by our model, and the rest is left unexplained. The part of this variation that is explained is

$$\text{ESS} = \sum_{\text{i}=1}^{n} (\hat{Y}_i - \bar{Y})^2. \tag{3.2}$$

We call this value the "Explained Sum of Squares," or ESS for short. It represents the explained behavior of Y about its mean.

Finally, as we have noted, the error term represents what is not explained by our model:

$$\text{RSS} = \sum_{\text{i}=1}^{n} \boldsymbol{e}_i^2. \tag{3.3}$$

We call this sum the "Residual Sum of Squares," or RSS. Using these measures, we can represent the behavior of Y in the following way:

$$\text{TSS} = \text{ESS} + \text{RSS}. \tag{3.4}$$

That is, the behavior of Y (TSS) can be broken up into two pieces: that which is explained by the model (ESS) and that which is unexplained (RSS). The R^2 is thus defined as the proportion of TSS explained by the model, or

$$R^2 = \frac{\text{ESS}}{\text{TSS}}. \tag{3.5a}$$

As an illustration of how to use the R^2, we can return to our second example, which looks at presidential voting. Looking at Table 2.2, we see the reported predicted values (i.e., the $\hat{Y}_i$ values). In Table A3 in Appendix A, we have the actual values for the percentage of two-party votes (i.e., the Y_i values) as well as the mean (i.e., the value for $\bar{Y}$). Thus, we have all the ingredients needed for calculating the R^2. Plugging these values into Equations 3.1 and 3.2, we find that TSS is 1023.61 and ESS is 539.52. Using Equation 3.5a, we have

$$R^2 = \frac{539.52}{1023.61} = 0.527. \tag{3.5b}$$

The result in Equation 3.5b means that approximately 53% of the variation in *Y* is explained by our model. This, in turn, tells us that about 47% of the behavior of *Y* is not explained by our model. These results suggest that our model is moderately successful, overall, in explaining the variation in the percentage of two-party votes received by incumbent party candidates.[4]

Virtually all computer packages that are capable of performing OLS regression will calculate the R^2. For example, Table 3.1 shows the OLS output generated by Excel for our baseball model.

As we can see, Table 3.1 provides a great deal of information.[5] We can see under the heading "Coefficients" the values for our intercept term, a (0.765), and the slope term, *b* (0.656), shown in Equation 2.5a. At the top of the table, we see the heading "Regression Statistics," below which we find "*R* Square," which is reported as 0.345. This is our measure of the "goodness of fit" discussed above. The interpretation of this number is that about 34.5% of the behavior (i.e., variation) of baseball salaries is explained by our model. This means that about two thirds of the variation in baseball salaries is left unexplained. This somewhat poor result should not be too surprising because, as was pointed out in Chapter 1, the two-variable model is too simplistic in most cases, and a more complex model is warranted.

Also shown in Table 3.1 is a section labeled "**ANOVA**," which stands for "*AN*alysis *O*f *VA*riance." Focusing on this section for a moment, we see a column headed "SS." This is short for "sum of squares" and is a breakdown of the variation of *Y* (baseball salaries) into its separate pieces as described in Equation 3.4.[6] The first entry for "Regression," 328.555, is our ESS in Equation 3.4. The second entry, labeled "Residual" (623.445), is our RSS in

[4] It should be pointed out that there is no benchmark R^2 value that needs to be achieved before we declare a model to be successful. There are some areas of research where an R^2 of 0.527 would be considered quite good (e.g., models of wage determination in industries other than MLB), whereas in other areas of research, it would be considered a weak result (e.g., models that forecast national income). These differences arise for a variety of reasons. It may be due to differences in the availability and quality of data, or it may simply be the case that some relationships naturally have a larger random component (i.e., u_i) than others.

[5] The output shown, as in earlier examples, has been edited to include only the relevant information needed for the present discussion. In addition, values were rounded to three decimal places.

[6] The column headed "MS" is simply the calculated mean of the sum of squares values. The MS values are not of particular use for us in this book.

Table 3.1

SUMMARY OUTPUT				
Regression Statistics				
Multiple *R*	0.587			
R Square	0.345			
Adjusted *R* Square	0.323			
Standard Error	4.559			
Observations	32			
ANOVA				
	df	SS	MS	
Regression	1	328.555	328.555	
Residual	30	623.445	20.782	
Total	31	952.000		
	Coefficients	Standard Error	*t* Stat	*P* value
Intercept	0.765	1.671	0.458	0.650
YEARS	0.656	0.165	3.976	0.000

Equation 3.4. And finally, “Total,” given as 952.000, is our TSS (= ESS + RSS) discussed earlier. Thus, using our definition of the R^2 in Equation 3.5a, we have

$$R^2 = \frac{328.555}{952.000} = 0.345, \tag{3.5c}$$

which is indeed our reported “*R* square” in Table 3.1.[7]

Several other measures are shown in Table 3.1 under the heading of “Regression Statistics,” and we can discuss them briefly. The **standard error**, shown as approximately 4.559, is the positive square root of the variance of the errors. It is essentially a measure of the typical size of the error (our e_i) in prediction. “Observations” simply tells us the sample size (32 in this case) used in the regression.

[7] “Multiple *R*,” reported as 0.587, is simply the square root of the R^2 and is the absolute value of the correlation between *Y* and *X*. This statistic is not often used in our assessment of the model’s overall performance.

Sample Results and Population Parameters

In addition to judging the overall performance of our model, we can consider the separate performance of estimated parameters *a* and *b*. It was pointed out in Chapter 2 that we typically are working with samples of data, not the population. Thus, we collect a sample and use it to calculate OLS values for *a* and *b*; these values then define a sample regression line. The hope is that our sample values of *a* and *b* are a good representation of the true, but unknown, values of the population parameters α and β shown in Equation 1.2b.

Suppose we replaced our original sample and collected a new one. This new sample could then be used to calculate OLS values for *a* and *b*. Obviously, because this second sample would not likely be identical to the previous one, these values for *a* and *b* would likely be different from those of the first sample. Different samples have different information contained in them; thus, as samples change, so will the OLS-calculated values for *a* and *b*. In other words, these sample values for *a* and *b* are random variables that "bounce around" from sample to sample and we can thus study their behavior. To illustrate the fact that *a* and *b* typically change from sample to sample, we can return to our baseball example. Suppose we draw two subsamples of size 16 from our 32 observations. We can then calculate OLS values for *a* and *b* for each subsample. For observations 1 through 16 (rounding to three decimal places), we have

$$a = 1.177, \qquad b = 0.680,$$

and for observations 17 through 32, we have

$$a = 0.352, \qquad b = 0.631.$$

As we can see, we have substantially different values for these estimated parameters from one sample to the next.[8]

Understanding that the OLS-calculated values for *a* and *b* are, in fact, random variables is important because we use them as estimates of the population's α and β. If the calculated values of *a* and *b* change a lot from sample to sample, then we would not have much confidence in any single sample's results being representative of α and β. On the other hand, if the OLS estimates of *a* and *b* differ only slightly from one sample to another, then we can be fairly confident that a single sample's results are a good representation of the population's parameters. The measures used to judge the

[8] The reader is invited to verify these results.

reliability of *a* and *b* as estimates of their population counterparts are the standard error of the coefficient estimates.[9] Essentially, they are measures of how *a* and *b* bounce around from sample to sample. The *smaller* the standard error, the *more reliable* are the sample values of *a* and *b* as estimates for α and β. In our baseball example, we can see in Table 3.1 that the column headed "Standard Error" for the "Intercept" (i.e., *a*) is shown as 1.671 and for "*YEARS*" (i.e., *b*) we have 0.165, after rounding to three decimal places. We need to keep in mind two things with these standard errors. First, they are in the same units as the dependent variable (millions of dollars in our case). Second, it is their size *relative* to the value of the estimated coefficient that is important for us. That is, consider the value for *b*, which is shown in Table 3.1 to be 0.656. Comparing this number to its standard error of 0.165, we see that the coefficient is nearly four times as large as the standard error. In other words, the sample value for *b* appears to be a fairly reliable estimate of the population's β.[10]

Given the values of the standard errors of our *a* and *b*, we can use them to test various ideas or "hypotheses." One of the most common tests is the "zero hypothesis" test. This test, in some sense, looks into the "soundness" of our model. Recall in Chapter 1 where we introduced the four examples. In each case, we specified a dependent (Y) variable and an independent (X) variable that we believed (or "hypothesized") should explain at least part of the behavior of the dependent variable. Take, for example, our model of voting patterns in presidential elections. In building this model, we hypothesized for the population data that the percentage of two-party votes received by the incumbent party candidate is determined, in part, by the economy's health as measured by the growth rate. This relationship between votes and growth is not an established fact; rather, it is a theory or hypothesis that we put forth. It could be wrong; there may be no relationship between votes and growth. Part of our task in regression analysis is to test this hypothesis. We do this by collecting a sample, estimating the sample regression function, and then analyzing our results to see whether the hypothesis is supported. More specifically, consider our model shown in Equation 1.6, except now for the population of data: $Y_t = \alpha + \beta X_t + u_t$.

If economic growth (i.e., variable X_t) is an important factor in explaining voting patterns (i.e., Y_t), then the population's value for β should be positive.

[9] See Gujarati (2003) for details on how the standard errors for the OLS parameter estimates are calculated.

[10] Generally speaking, if the absolute value of the estimated parameter (e.g., *a* or *b*) is twice (or more) the size of the associated standard error, then the estimated parameter is considered to be a fairly reliable estimate of the population parameter (e.g., α or β).

On the other hand, suppose that the economic growth rate has nothing to do with the way people vote; in that case, β would be zero. If β is truly zero, then βX_t would be zero, meaning that X_t has no impact on Y_t.

Whether or not β is zero is a hypothesis made for our population of data. Stating the hypothesis a bit more formally, we can write

$$H_0\text{: } \beta = 0$$

$$H_1\text{: } \beta \neq 0,$$

where H_0 is called the **null hypothesis** and H_1 is the **alternative hypothesis**. If we find that H_0 is true, then our hypothesis that economic growth is an important factor in explaining voting patterns would be false. On the other hand, if we find that H_1 is true, then our hypothesis is correct. Unfortunately, we typically cannot prove that either H_0 or H_1 is true. This is because we typically do not have the population data and are working only with a sample. Because our sample result for *b* is not a perfect predictor for β, our sample results can only support or fail to support our hypotheses over β.[11]

In order to see how we can use our sample regression results to test the above hypothesis, we can use the information in Table 3.2, which contains the OLS output (produced by Excel) for our voting model.[12]

We see in Table 3.2 that the estimated values for *a* and *b* are, as we saw earlier, 51.065 and 0.880, respectively. We can also see that the standard errors for these values of *a* and *b* are approximately 1.035 and 0.182, respectively. If we next consider the column headed "*t* stat" (short for ***t* statistic**), we see the rounded value of 49.339 for the intercept, *a*, and 4.838 for the coefficient to *GROWTH*, or *b*. The basis of the *t* statistic actually stems from the normality assumption that was introduced in Chapter 2 in CLRM Assumption 7. Although the discussion of the use of the *t* statistic is somewhat advanced, it boils down to this: If the population errors, u_i, are normally distributed, then it can be shown that the sample estimates for *a* and *b* follow a *t* distribution.[13] Given this result, then we can use the *t* distribution to test hypotheses over the population parameters α and β with our sample estimates. In other words, the

[11] It should be understood that β is a fixed parameter, not a random variable. The value for *b* from our sample regression, on the other hand, is not a fixed parameter but is a random variable that we use to estimate β.

[12] Table 3.2 is similar to Table 2.2, except that the prediction errors and predicted values are omitted and other relevant statistics are now included.

[13] See Gujarati (2003) for the proof of this result.

Table 3.2

SUMMARY OUTPUT				
Regression Statistics				
Multiple *R*	0.726			
R Square	0.527			
Adjusted *R* Square	0.505			
Standard Error	4.801			
Observations	23			
ANOVA				
	df	SS	MS	
Regression	1	539.524	539.524	
Residual	21	484.088	23.052	
Total	22	1023.612		
	Coefficients	Standard Error	*t* stat	*P* value
Intercept	51.065	1.035	49.339	0.000
GROWTH	0.880	0.182	4.838	0.000

normality assumption introduced in Chapter 2 now justifies our use of the *t* stat here.

As for the value for each *t* stat shown, it is calculated by dividing a coefficient by its associated standard error. That is, for the intercept, if we divide 51.065 by 1.035, we get 49.339, which is the *t* stat value shown (allowing for rounding differences). Similarly, for the coefficient to *GROWTH* (i.e., our *b*), we have 0.880, which, if divided by its standard error of 0.182, gives us 4.838, the reported *t* stat for *GROWTH*. As noted earlier, the size of the standard error relative to the estimated coefficient is what is important for us. This is exactly what the reported *t* stat shows: the coefficient divided by its own standard error. In the case of the intercept, our *t* stat of 49.339 tells us that our calculated value of a is about 49 times larger than its standard error. In other words, the OLS value for a is a reliable estimate of the population's α because it bounces around very little from sample to sample. As for *b*, this coefficient is approximately 4.8 times larger than its own standard error, indicating that this sample value for *b* is a very reliable estimate for the population's parameter β. Exactly how reliable are these estimates? The answer to this question is our next task.

The column next to the *t* stat column in Table 3.2 is headed "***P* value.**" The numbers shown here for *a* and *b* are calculated using the *t* stat.[14] These numbers represent *probabilities,* based on the associated *t* stat, that we can use to test the zero hypotheses discussed earlier. The interpretation of these numbers is a little tricky. They represent what the probability would be of finding the values of *a* and *b* shown if, in fact, *for the population,* the null hypothesis was true. In other words, it sets up a straw man that we hope we can knock down, and this straw man says that, for the population, *X* has no relationship to *Y*.

In order to make this clear, let's focus on the value for *b* shown in Table 3.2. This coefficient to *GROWTH* is estimated to be 0.880, based on our sample OLS results. This number, as we have discussed, is only an estimate of the true value of the population's β. Furthermore, we have learned that this result for *b* will likely differ from sample to sample (i.e., it is a random variable). This being the case, the value of 0.880 could be quite different from β. We can therefore consider the following question: Could it be the case that the true value for β is, in fact, zero, and that our value of 0.880 was merely a result of our sampling? This is the question that the *P* value addresses: What is the likelihood of getting a sample value of 0.880 when, in fact, it is true that β is zero? In our case, the *P* value for *b* is shown as 0.000 after rounding to three decimal places. If we allow for a greater number of decimal places, say six, we would find this *P* value to be 0.000088.[15] That is, the probability of getting the value of 0.880 for *b* when it is true that β is zero is about 0.000088 (or about a 0.0088% chance).[16] This is quite small,

[14] The probabilities shown under the *P* value column are calculated by plugging the given *t* stat into the *t* probability distribution function. This, again, is a complicated matter that goes beyond the scope of this book. Those interested in learning more about the specifics of probability distribution functions are referred to Gujarati (2003) or Greene (2003).

[15] In fact, the Excel output produces the following *P* value for the coefficient to *GROWTH*: 8.79747E-05. This number is expressed in scientific notation and could be rewritten as 8.79747×10^{-5} or, in nonscientific notation, as 0.0000879747.

[16] This interpretation is not strictly correct. The hypothesis test we are conducting is whether or not β is statistically different from zero. To reject this null hypothesis, we need to get a sample result for β that is sufficiently different from zero, that is, sufficiently larger *or* sufficiently smaller than zero. In our case, our sample result was 0.880, but if our sample gave us a value of –0.880, this too would be sufficiently different from zero (in this case, smaller than zero) to reject our null hypothesis. Thus a more accurate interpretation of the *P* value in this example would be: The probability of obtaining a sample value of 0.880 or greater *in absolute value terms*, when the true population's value is zero, is approximately 0.000088.

indicating that the hypothesis that β is zero is probably not true. In other words, we would likely reject H_0: $\beta = 0$ in favor of H_1: $\beta \neq 0$.

We can also perform a similar test for the intercept term, α. That is, we can consider the hypothesis that the population's intercept is really equal to zero. Putting it formally, we would have

$$H_0: \alpha = 0$$

$$H_1: \alpha \neq 0.$$

In order to test this hypothesis, we need only consider the *P* value for a, which is given as approximately 0.000. Or, again, allowing for a greater number of decimal places, the Excel program reports 3.3542E-23. This is in scientific notation, which means we move the decimal 23 places to the left. Obviously, this is an extremely small number. In terms of our hypothesis test, this means that the probability of obtaining a value for a of 51.065 from our sample when the value for α is truly zero is 3.3542E-23 (or a 3.3542E-21% chance). In other words, it means we can be quite confident that the true population's intercept α is not zero.

In the above example, we found very small *P* values, which indicated that we can reject H_0 for both the intercept and slope terms with great confidence. Suppose, however, that we were not able to be so confident. That is, suppose the *P* value for *b* turned out to be 0.20. This would mean that the chance of getting the value of *b* we found (or greater, in absolute terms), when in fact the population's β is zero, is about 20%. Would we be sufficiently confident that β is not zero with this result? In other words, how large of a *P* value are we willing to accept before concluding that we cannot reject H_0? If the *P* value is 0.20 and we declare that we reject H_0, then this means we have a 20% chance that our declaration is wrong.[17] How prudent we are in rejecting the null hypothesis (i.e., accepting the estimated relationship) may vary with circumstances. For example, in a model exploring the usage of a particular drug and the possibility of its use causing birth defects, we would want to be quite sure that a real relationship exists. It is standard practice to reject the zero hypothesis when the associated *P* values are 0.05 or smaller (in some cases, *P* values of 0.10 or smaller are used).

In order to solidify the concept of hypothesis testing, we can apply these tools to our third example. Recall that this model, which is shown in Equation 1.7, hypothesizes that abortion rates across the 50 states differ, in

[17] If we reject a hypothesis when it is actually true, we are committing what is called a "Type I error." (Accepting a hypothesis when it is actually false is called a "Type II error.") The *P* value thus represents the probability of committing a Type I error.

part, because the moral views on abortion differ across the states. We use the variable *RELIGION*, which shows the percentage of the population that is Catholic, Southern Baptist, Evangelical, or Mormon, to capture the moral differences. As *RELIGION* (X_i) increases from one state to another, we expect that the abortion rate (Y_i) would be smaller, other things being equal. As in the previous example, this is a hypothesis that we put forth. It could be wrong, meaning that *RELIGION* tells us nothing about abortion rates. That is, it may be the case that β is zero for the population. We can test this hypothesis by evaluating the *P* value from the OLS regression. Table 3.3 shows the output from our earlier regression, but with more details in this case.

The three panels of output shown in Table 3.3 were produced by the program SPSS. The output provided is similar to that produced by the Excel program. We can see in the first panel, titled "Model Summary," that the *R* squared is reported as 0.016. This result tells us that approximately 1.6% of the behavior of abortion rates (Y_t) is explained by our model. This is

Table 3.3

Model Summary

Model	*R*	*R* Square	Adjusted *R* Square	Std. Error of the Estimate
1	.125[a]	.016	−.005	10.0829

a. Predictors: (Constant), *RELIGION*

ANOVA[b]

Model		Sum of Squares	df	Mean Square
1	Regression	77.691	1	77.691
	Residual	4879.935	48	101.665
	Total	4957.626	49	

b. Dependent Variable: ABORTION RATE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	*t*	Sig.
1	(Constant)	23.825	3.979		5.988	.000
	RELIGION	−.099	.114	−.125	−.874	.386

a. Dependent Variable: ABORTION RATE

clearly a very poor result, because it means that approximately 98.4% of the behavior (i.e., variation) of our dependent variable is not explained. This result is mirrored in the second panel, titled "ANOVA," which shows that the residual sum of squares, RSS (reported as 4879.935), makes up nearly all of the total sum of squares, TSS (reported as 4957.626). In sum, the model, as it stands, fits the data very poorly.

Turning to the separate components of the model, we find similarly weak results. In the panel titled "Coefficients," we find the estimate of a to be 23.825 with a standard error of 3.979, and the estimate of *b* to be −0.099 with a standard error of 0.114. In the case of the constant term a, dividing the coefficient by its standard error (i.e., 23.825/3.979) gives us a reported "*t*" (which is the same as Excel's "*t* stat") of 5.988. This *t* then translates into a "Sig." (short for **significance level**, which is the equivalent of Excel's *P* value) of 0.000. This result tells us that we can reject the hypothesis that the constant term for the population, α, is truly zero with a very high degree of confidence.[18]

As for the coefficient to *RELIGION* (i.e., *b*), the results are quite different. The coefficient has the negative sign that was hypothesized, supporting the idea that as *RELIGION* increases from state to state, abortion rates tend to be smaller, other things being equal. However, the value for *b* appears not to be very different from zero. Again, we can ask the question: Is it true that the population value for β is really zero and that our result for *b* of −0.099 is from a sample that does not perfectly reflect the population? In other words, is –0.099 *statistically* different from zero? The fact that the magnitude of our *b* is small is not enough to make any conclusions about whether it is not statistically different from zero. We must compare the value for *b* to its standard error. The reported standard error of *b* is *larger* than (the absolute value of) *b*. In other words, for this model, the random variable *b* tends to bounce around a great deal, and as such the computed value is not a very reliable predictor of β. We can see this clearly by observing the Sig. value for *b*, shown as 0.386. What this number means is that there is a 38.6% chance of finding a sample value of 0.099 or larger (in absolute terms) for *b* when, in fact, the population's β is truly zero.[19]

[18] The Sig. value of 0.000 represents a rounded figure; it is not perfectly equal to zero, just very close. The fact that the intercept term is significantly different from zero, however, is not of great importance to us because the interpretation of the intercept in this case is not very meaningful.

[19] A word on terminology: In determining whether we reject or do not reject the null hypothesis, the *P* value (or Sig.) tells us at what probability the coefficient is "significantly" different from zero. For example, a *P* value of 0.10 means that we can reject the hypothesis that the associated coefficient is zero at the 10% "significance level." Equivalently, we can say that the null hypothesis is rejected at the 90% "confidence level." The terms *significance level* and *confidence level* are used interchangeably, and the confidence level is simply 100 minus the significance level.

In other words, we cannot confidently rule out the possibility that *RELIGION*, on its own, has nothing to say about state abortion rates.

On some occasions, we must work directly with the *t* statistic in order to test hypotheses. This may occur, for example, if the software program with which we are working does not produce a *P* value (or significance level) for estimated coefficients.[20] Alternatively, we may wish to test a hypothesis other than the "zero hypothesis" test we have been performing. In both cases, we need to do some work by hand to perform a hypothesis test for our estimated coefficients. In order to see how we can work directly with the *t* statistic, we can return to our crime example. Table 3.4 reports the same results as Table 2.4 except we now include more detailed output.

Using the data in Table 3.4, we can write down the estimated equation, which is the same as we saw in Equation 2.9b but now including the standard error ("*se*" for short) for the intercept and coefficient to $UNEM_i$ shown in parentheses below the estimated coefficients[21]:

$$\widehat{CRIME}_i = 51.772 + 10.218\ (UNEM_i)$$

$$se = (5.698) \quad (1.829)$$

Suppose we wanted to perform a zero hypothesis test for the estimated coefficient to *UNEM*, *b*, but we did not have a *P* value to use. In order to do so, we can use the standard errors provided to compute the *t* statistic, and then we can compare it to a table of *t* values that is provided in Appendix C of this book. Before describing how to use the table of *t* values, however, we can explore the intuition behind this procedure.

Recall that the value of *b* is a sample's estimate of the population's coefficient to *UNEM*, β. The standard error of *b* is a measure of how precise our estimate is: the larger the standard error, the less precise our estimate. We noted that the "*t* stat" reported by Excel (or the "*t*" reported by SPSS) is calculated by dividing the estimated coefficient by its standard error. That is, for the coefficient to *UNEM*, we have

$$t\text{ statistic} = \frac{b}{se(b)}. \qquad (3.6a)$$

[20] Some older software programs simply provided the standard error for each coefficient estimated, and the researcher then had to perform calculations and hypothesis tests by hand. Most software programs now routinely generate *t* statistics and *P* values (or significance levels).

[21] Putting the standard errors in parentheses below their respective estimated coefficients is a common way to report regression results.

Table 3.4

Model Summary

Model	*R*	*R* Square	Adjusted *R* Square	Std. Error of the Estimate
1	.662[a]	.438	.424	13.610087

a. Predictors: (Constant), *UNEM*

ANOVA[b]

Model		Sum of Squares	df	Mean Square	*F*	Sig.
1	Regression	5783.392	1	5783.392	31.222	.000[a]
	Residual	7409.379	40	185.234		
	Total	13192.77	41			

a. Predictors: (Constant), *UNEM*

b. Dependent Variable: CRIME

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		*B*	Std. Error	Beta	*t*	Sig.
1	(Constant)	51.772	5.698		9.085	.000
	UNEM	10.218	1.829	.662	5.588	.000

a. Dependent Variable: CRIME

Focusing on Equation 3.6a for the moment, we see that for any given value for *b*, the larger the se(*b*), the smaller the *t* statistic. Therefore, the smaller the *t* statistic, the less reliable our value of *b* is as an estimate of the population's coefficient, β. In other words, a "small" absolute value of the *t* statistic means that our value of *b* bounces around a lot from sample to sample. This, in turn, means that the population's value for β may truly be zero and that our sample estimate of the coefficient was simply off the mark. Putting this formally, we can state the following two hypotheses:

$$H_0\text{: } \beta = 0$$

$$H_1\text{: } \beta \neq 0.$$

In this case, the "smaller" the absolute value of the *t* statistic, the more likely it is that we would not reject H_0 (i.e., that the population's regression line has a zero coefficient to *UNEM*). On the other hand, the "larger" the absolute value of the *t* statistic, the more likely it is that we would reject H_0 in favor of H_1 (i.e., that the population's regression line has a nonzero coefficient for *UNEM*). The question that arises is, how "large" does the *t* statistic have to be to reject H_0? The answer to this question depends on how confident we want to be in our test's result. It is standard practice that researchers settle on 90% or 95% confidence levels, or equivalently, 10% or 5% significance levels (see footnote 19). Thus, if we choose a significance level of 10% as our cutoff, we can then determine if our *t* statistic is large enough to reject H_0 for this level. To carry out this test, we can refer to Table C1 in Appendix C. In the first column of this table, we see the heading "df." This is short for **degrees of freedom**, which are calculated by taking our sample size and subtracting from it the number of parameters we have estimated.[22] Thus, for our 2004 crime example, we have a sample size of 42 police force areas and we are estimating two parameters (*a* and *b*), leaving us with 40 degrees of freedom.[23] Referring to Table C1, we see that the first column is headed "df" for degrees of freedom. For this test, we go down to the row for 40 degrees of freedom.[24] Across the top of Table C1, we see the heading "Confidence Level." Below this. we see "Probability," which, if multiplied by 100, is our significance level. Staying with our chosen 10% significance level, this means we go over to the column headed by 0.1 (i.e., 10%). We see, then, that for 40 degrees of freedom and a significance level of 0.10, we have the number 1.684 in Table C1. This number is a *t* value, and it represents the minimum value

[22] The meaning of the degrees of freedom is somewhat difficult to explain. The easiest way to think about the degrees of freedom is that they measure how much useable information we have left over from our sample after already performing certain tasks. That is, when we have a sample of data, this data set contains a finite amount of information. If we use part of this information to perform certain calculations, such as estimating *a* and *b* in a sample regression function, then we have less information left over to carry out other tasks, such as hypothesis tests. Thus, the degrees of freedom keep track of how much information we have left for such tasks.

[23] Note that both Excel and SPSS routinely report the degrees of freedom. Both programs provide several values for the degrees of freedom, shown as "df," in the ANOVA portion of their output. The relevant value for the purpose of performing *t* tests are the ones reported for the "Residual." Thus, we see in Table 3.4 a value of 40. In the case of our abortion model, with a sample size of 50, we have degrees of freedom equal to 48 (see Table 3.3).

[24] Other, more extensive *t* tables have a greater number of values for degrees of freedom.

that the absolute value of our *t* statistic must achieve before we can reject H_0.[25] That is, if

$$|t \text{ statistic}| \geq t \text{ value},$$

then we can reject H_0, meaning that the coefficient to *UNEM* is statistically different from zero at the 10% significance level.

Returning to our crime example, we have

$$t \text{ statistic} = \frac{10.218}{1.829} = 5.587,$$

which is the same value reported for the *t* statistic in Table 3.4 (allowing for some minor difference due to rounding) for the coefficient to *UNEM*. Because this *t* statistic is greater than 1.684, the *t* value, we can reject H_0 at the 10% significance (90% confidence) level. In fact, we can see that we can achieve even a smaller significance (or higher degree of confidence) with our given *t* statistic. Our *t* statistic is larger than the *t* value of 0.05 (shown as 2.021 in Table C1) and is also larger than the *t* value for 0.01 (shown as 2.704). Thus, we can reject H_0 at a significance level that is smaller than 1%. What is the exact significance level we can achieve? This is, in fact, what the Sig. value tells us. Referring back to Table 3.4, the Sig. value is shown as 0.000, which is rounded to three decimal places. Carrying the Sig. value out to 6 decimal places, we have 0.000002 as our significance level. This Sig. value tells us that we can reject H_0 at less than the 0.0002% level of significance. This example illustrates the linkage between the *t* statistic reported by the program and the associated Sig. value. Furthermore, it provides us with strong evidence that unemployment is truly an important determining factor of British police force area crime rates in 2004.

We can also use *t* statistics to perform hypothesis tests other than the zero hypothesis. For example, we saw in our voting regression (see Table 3.2)

[25] If the value for *a* or *b* happens to be negative, then the *t* statistic will be negative (e.g., the value for *b* in our abortion regression shown in Table 3.3). The *t* tables, however, show only positive values. This is because the *t* distribution is symmetric and centered around zero, so if we compute a *t* statistic and find it is negative, we can take the absolute value of this number and compare it to the positive *t* values shown in Table C1. For a more detailed discussion of the *t* distribution, see Gujarati (2003).

that the OLS estimation of the slope term, *b*, yielded a value of approximately 0.880 with a standard error of about 0.182 (rounding to three decimal places). The "*t* stat" and the associated "*P* value" show that the population's β is different from zero at a very small significance level (or high confidence level). We can, however, ask the following question: Is the population's β different from 1? The value of 0.880 is not too far off from 1, and given the fact that our value for *b* is from a sample, it seems possible that if we had the population data and calculated the population regression function, we could perhaps end up with β equal to 1. Recall, however, that we cannot simply consider the size of *b*; we must also consider its size relative to its standard error. Thus, a more formal test is needed. We can state our hypothesis more formally as

$$H_0\colon \beta = 1$$

$$H_1\colon \beta \neq 1.$$

We can perform this test using the *t* statistic in a similar way as we did above. In this case, however, we need to recalculate the *t* statistic to make it conform to our new hypothesized value of β. We calculate this new *t* statistic using the following equation:

$$t\text{ statistic} = \frac{b - h}{se(b)}, \tag{3.6b}$$

where *b* is the OLS sample result for the slope term and *h* is our hypothesized value for β.[26] Using the values from our voting regression, we have

$$t\text{ statistic} = \frac{0.880 - 1}{0.182} = -0.659, \tag{3.6c}$$

and taking the absolute value for this *t* statistic, we have 0.659. Turning to Table C1 and locating the row for 21 degrees of freedom (23 observations

[26] This equation is very similar to that shown in Equation 3.6a, except that we are subtracting *h* in the numerator. However, this is only an apparent difference. Recall in Equation 3.6a that we were calculating a *t* statistic for use in testing a hypothesized value of zero for our population parameter β. In that case, then, *h* was zero, and so *b* minus *h* would simply be equal to *b*.

minus two estimated parameters), we see that for a significance of level of 10% (or confidence level of 90%), the *t* value is 1.721. Comparing the absolute value of our *t* statistic to this *t* value, we see that it is less than the *t* value, and so we cannot reject H_0, that β is truly 1 for the population at the 10% level of significance. In fact, we see that the absolute value of our calculated *t* statistic is smaller than all the values reported for 21 degrees of freedom. The bottom line is that we cannot confidently reject the hypothesis that β is equal to 1.

Summing Up

In building regression models, the researcher is responsible for specifying what factors are important in explaining the behavior of a dependent variable. Whether or not the researcher has built a sound model depends on a number of things. Ultimately, though, the model must have some logic to it. In our four examples, we put forth hypotheses about how *Y* can be explained by the related *X* variable. Using samples of data, we then calculated the OLS regression for each model and asked the following questions: How well does the model fit the data as a whole? And how do the separate components of the model (i.e., *a* and *b*) perform? We have seen that the first question can be answered by using the R^2 value. If the R^2 is "large" (i.e., close to 1), then our model works well, and if it is "small" (i.e., close to 0), then our model performs poorly. What value of the R^2 is large enough so that we can claim a good fit is subjective.[27]

With regard to the second question, we have seen that an estimated coefficient's standard error can be used to determine if the coefficient is reliably different from zero. Using the *P* value (or Sig. in the case of SPSS), we have determined that the smaller its value, the more confident we are that the variable included in the model is truly important in explaining the behavior of the dependent variable. We have also seen that we can test other hypothesized values for our parameters by calculating *t* statistics and comparing them to *t* values provided in Table C1 in Appendix C.

[27] It should be noted that it is possible for a model to produce a relatively large R^2 even though each separate estimated coefficient fails to achieve a sufficient level of significance. It is also possible to achieve a low R^2 when each estimated coefficient is highly significant.

PROBLEMS

3.1 Use Excel (or another program) and the data shown in Table A6 in Appendix A to calculate the OLS estimation of the following model:

$$Y_i = \alpha + \beta X_i + u_i$$

where: Y_i is hourly wage
X_i is years of education.

Interpret the estimated values for α and β. What is the R^2 for this regression? What is its interpretation?

3.2 Use the regression output from Problem 3.1 to perform the following hypothesis tests:

a. H_0: $\alpha = 0$, H_1: $\alpha \neq 0$, at the 5% significance (or 95% confidence) level.
b. H_0: $\beta = 0$, H_1: $\beta \neq 0$, at the 1% significance (or 99% confidence) level.

3.3 Use the output shown in Table 3.4 and the *t* table presented in Appendix C to test the following hypothesis:

H_0: $\beta = 9$, H_1: $\beta \neq 9$, at the 5% significance (or 95% confidence) level. [Hint: You should use Equation 3.6b to conduct this test.]

3.4 Using SPSS or Excel (or an equivalent program), perform an OLS regression with GPA as the dependent variable and SAT as the independent variable and conduct the following hypothesis test:

H_0: $\alpha = 0.8$, H_1: $\alpha \neq 0.8$, at the 5% significance (or 95% confidence) level. [Hint: You should use Equation 3.6b to conduct this test.]