# 2

# The Least-Squares Estimation Method

## Fitting Lines to Data

In the various examples discussed in the previous chapter, lines were drawn in such a way as to best fit the data at hand. The question arises as to how we find the equation to such a line. This is the point of linear regression analysis: fitting lines to data. We can consider a number of approaches. For example, we could consider simply using a ruler and drawing a line that seems to fit the data best. This method, however, is not advisable because it is not a precise approach, and in some cases, the scatter of points does not suggest to the naked eye an obvious location for the line. A more systematic approach is needed.

One possibility would be to find a line that would minimize the sum of the error terms. This approach, however, is flawed. To see this, suppose we have a very small sample of just four data points, which are plotted in Figure 2.1.

The line shown seems to fit the data well using the criterion that the sum of the errors is minimized. In fact, if we calculate $e_1 + e_2 + e_3 + e_4$, it would come to approximately zero as the small positive errors ($e_2$ and $e_4$) would cancel with the small negative errors ($e_1$ and $e_3$), indicating a good fit. Unfortunately, under this criterion, the line shown in Figure 2.2 would serve just as well.

As in the previous case, the sum of the errors here is also approximately zero as the small positive error cancels with the small negative error ($e_3$ with $e_2$) and the large positive error cancels with the large negative error ($e_4$ with $e_1$). Thus, we have two very different lines that meet our criterion of minimizing the sum of the

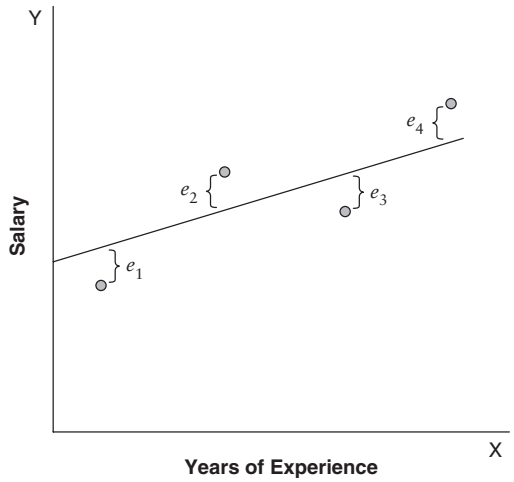


**Figure 2.1**

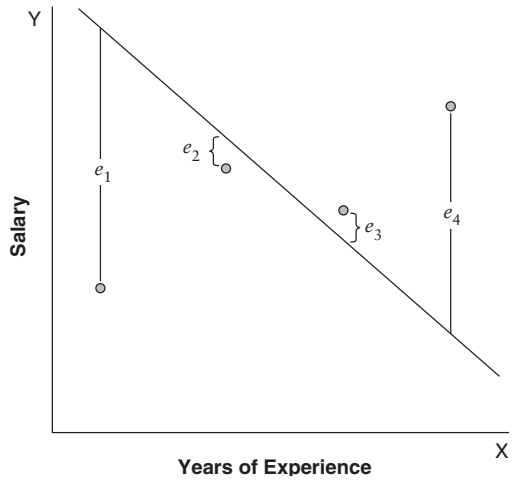


**Figure 2.2**

errors equally as well. In fact, there are any number of such lines that, when drawn, would give us a zero sum of errors. Obviously, this criterion will not do.

The problem with the last method considered is the cancellation of positive errors with negative errors. A way of avoiding this problem would be to find a line such that the sum of the *squared* errors is minimized.[1] That is,

[1] We could consider minimizing the sum of the absolute value of the errors, but such a method is computationally difficult.

our task is to find a line determined by $a$ and $b$ such that the sum of the squared errors, $e_1^2 + e_2^2 + e_3^2 + e_4^2$, is as small as possible. This, in fact, is called the method of least-squares, sometimes called **ordinary least-squares** (**OLS**) so as to distinguish it from other specialized least-squares methods.[2] We can represent this task of minimizing the sum of squared errors mathematically by first noting that the error term $e_i$ can be rewritten using Equation 1.5 in the following way:

$$e_i = Y_i - (a + bX_i). \tag{2.1a}$$

Thus, the error is decomposed into the difference between the observed value $Y_i$ and the predicted value $(a + bX_i)$. Or, getting rid of the parentheses, we have

$$e_i = Y_i - a - bX_i. \tag{2.1b}$$

The OLS method finds a value for $a$ and $b$ that minimizes the sum of the squared errors; thus, we take the errors shown in Equation 2.1b, square them, and then take the sum over all observations in our sample (the sample size indicated by $n$). Doing this, we have

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2. \tag{2.2}$$

At this point, our task now becomes a calculus problem. We have an equation (Equation 2.2) that we want to minimize with respect to two parameters we can choose, $a$ and $b$. To carry out this task, we use differential calculus and take the partial derivative of Equation 2.2 with respect to $a$ and set it equal to zero, then do the same with respect to $b$. We then end up with two equations (the two partial derivatives) with two unknowns ($a$ and $b$). The last step is to solve this system of equations simultaneously for $a$ and $b$. Leaving out the details, what we end up with is the following formulas for $b$ and $a$[3]:

$$b = \frac{\sum_{i=1}^{n} (\mathrm{X}_i - \bar{\mathrm{X}})(\mathrm{Y}_i - \bar{\mathrm{Y}})}{\sum_{i=1}^{n} (\mathrm{X}_i - \bar{\mathrm{X}})^2} \tag{2.3a}$$

[2] There are other, advanced methods, such as "two-stage least-squares" or "weighted least-squares," that are used in certain circumstances. These methods are beyond the scope of this book. See, for example, Gujarati (2003) or Wooldridge (2006) for a discussion of these techniques and others.

[3] For the full details of solving of the OLS estimators, see Gujarati (2003).

and

$$a = \bar{Y} - b\bar{X}, \tag{2.3b}$$

where, in both Equations 2.3a and 2.3b, the bar above the variable stands for the mean of that variable (e.g., the sum of all $X_i$ divided by the sample size, $n$). Thus, given values for $X_i$ and $Y_i$, we can use these data to calculate $a$ and $b$.

As an illustration, we can go back to our baseball example. Using the data on salary ($Y_i$) and years of MLB experience ($X_i$) provided in Table A1 in Appendix A, we can construct Table 2.1, which has the components needed for computing $b$ and $a$.[4]

Viewing Table 2.1, we see that the sum at the bottom of the sixth column is the value needed for the numerator in Equation 2.3a. And the sum at the bottom of the fifth column is the value needed for the denominator. Plugging these values into our equation for $b$ (and rounding to three decimal places), we have the following:

$$b = \frac{500.851}{763.500} = 0.656. \tag{2.4a}$$

As for the intercept term, $a$, dividing the sums at the bottom of the first and second column by our sample size, $n = 32$, we find

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{284}{32} = 8.875$$

and

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{210.795}{32} = 6.587.$$

Using these values for the mean of $X$ and $Y$, along with the value for $b$ found in Equation 2.4a, we have for our intercept (again, rounding to three decimal places):

$$a = 6.587 - (0.656 \times 8.875) = 0.765. \tag{2.4b}$$

[4] Our analysis will consider only nonpitchers because pitchers are evaluated with very different statistical measures.

**Table 2.1**

| $X_i$ | $Y_i$ | $X_i-\bar{X}$ | $Y_i-\bar{Y}$ | $(X_i-\bar{X})^2$ | $(X_i-\bar{X})(Y_i-\bar{Y})$ |
|---|---|---|---|---|---|
| 10 | 13.600 | 1.125 | 7.013 | 1.266 | 7.890 |
| 12 | 10.600 | 3.125 | 4.013 | 9.766 | 12.541 |
| 18 | 4.000 | 9.125 | –2.587 | 83.266 | –23.606 |
| 9 | 3.750 | 0.125 | –2.837 | 0.016 | –0.355 |
| 4 | 0.800 | –4.875 | –5.787 | 23.766 | 28.212 |
| 8 | 9.500 | –0.875 | 2.913 | 0.766 | –2.549 |
| 3 | 0.354 | –5.875 | –6.233 | 34.516 | 36.617 |
| 5 | 3.333 | –3.875 | –3.254 | 15.016 | 12.608 |
| 3 | 0.425 | –5.875 | –6.162 | 34.516 | 36.202 |
| 10 | 6.000 | 1.125 | –0.587 | 1.266 | –0.660 |
| 5 | 4.200 | –3.875 | –2.387 | 15.016 | 9.250 |
| 17 | 10.465 | 8.125 | 3.878 | 66.016 | 31.506 |
| 10 | 13.500 | 1.125 | 6.913 | 1.266 | 7.777 |
| 7 | 4.000 | –1.875 | –2.587 | 3.516 | 4.851 |
| 11 | 20.600 | 2.125 | 14.013 | 4.516 | 29.778 |
| 13 | 12.333 | 4.125 | 5.746 | 17.016 | 23.704 |
| 9 | 12.000 | 0.125 | 5.413 | 0.016 | 0.677 |
| 2 | 0.340 | –6.875 | –6.247 | 47.266 | 42.948 |
| 6 | 0.850 | –2.875 | –5.737 | 8.266 | 16.494 |
| 14 | 9.000 | 5.125 | 2.413 | 26.266 | 12.367 |
| 4 | 1.000 | –4.875 | –5.587 | 23.766 | 27.237 |
| 6 | 0.700 | –2.875 | –5.887 | 8.266 | 16.925 |
| 5 | 0.690 | –3.875 | –5.897 | 15.016 | 22.851 |
| 15 | 10.616 | 6.125 | 4.029 | 37.516 | 24.680 |
| 6 | 5.000 | –2.875 | –1.587 | 8.266 | 4.563 |
| 18 | 10.756 | 9.125 | 4.169 | 83.266 | 38.044 |
| 5 | 12.500 | –3.875 | 5.913 | 15.016 | –22.913 |
| 11 | 11.000 | 2.125 | 4.413 | 4.516 | 9.378 |
| 2 | 0.400 | –6.875 | –6.187 | 47.266 | 42.536 |
| 15 | 14.167 | 6.125 | 7.580 | 37.516 | 46.425 |
| 17 | 3.640 | 8.125 | –2.947 | 66.016 | –23.944 |
| 4 | 0.675 | –4.875 | –5.912 | 23.766 | 28.821 |
| Sums: 284 | 210.795 | | | 763.500 | 500.851 |

Thus, putting it all together, we have for the OLS regression line

$$\hat{Y}_i = 0.765 + 0.656\, X_i, \tag{2.5a}$$

where the "hat" above $Y_i$ denotes the predicted salary for player *i*. As an alternative way of reporting results, we can replace *Y* and *X* with their variable names:

$$\widehat{SALARY}_i = 0.765 + 0.656\ (YEARS_i). \tag{2.5b}$$

Interpreting these results, the intercept term, 0.765, is the predicted salary (in millions of dollars) for a player who has no MLB experience (i.e., a rookie). The value of *b*, 0.656, represents the added salary that a player earns, on average, for each additional year he plays in the MLB. Or, an additional year of experience adds on average about $656,000 to salary, all else equal. Thus, a player who has 5 years of MLB experience is expected to earn

$$\widehat{SALARY}_i = 0.765 + 0.656\ (5) = 4.045. \tag{2.5c}$$

Comparing this predicted salary to MLB player David Eckstein, who has 5 years of experience (see Table A1), we see that his actual salary for the 2006 season was $3.333 million, which is $712,000 *less* than what his predicted salary would be according to Equation 2.5c. On the other hand, if we consider the salary for 5-year player Jay Gibbons, his actual salary was $4.2 million, or $155,000 *more* than predicted. The difference between these actual figures and the predicted ones is captured by the error term, $e_i$. Essentially, what we learn from this result is that years of experience may be important, but there are other factors in addition to experience that determine a player's salary.[5] Figure 2.3 shows the regression line (Equation 2.5a) plotted with the actual data for player salaries and years of experience. The fact that the plotted points do not fall precisely on the regression line illustrates this last point.

Applying the same least-squares method to our example for U.S. presidential elections, we can estimate the relationship between economic growth and presidential voting patterns. Using the data shown in Table A3 in Appendix A, we could plug the values for the dependent variable (the column of data labeled with *VOTES*) for $Y_t$ and the values for the independent variable (the column of data labeled with *GROWTH*) for $X_t$ into Equations 2.3a and 2.3b to determine the value for the slope, *b*, and the intercept, *a*, for the sample regression function shown in Equation 1.6. Although it was instructive

[5] Explaining part of the discrepancy between Eckstein's and Gibbons's salaries is the fact that Gibbons has a better career slugging average of 46.6 (through the 2005 season) compared to Eckstein's career slugging average of 36.2.

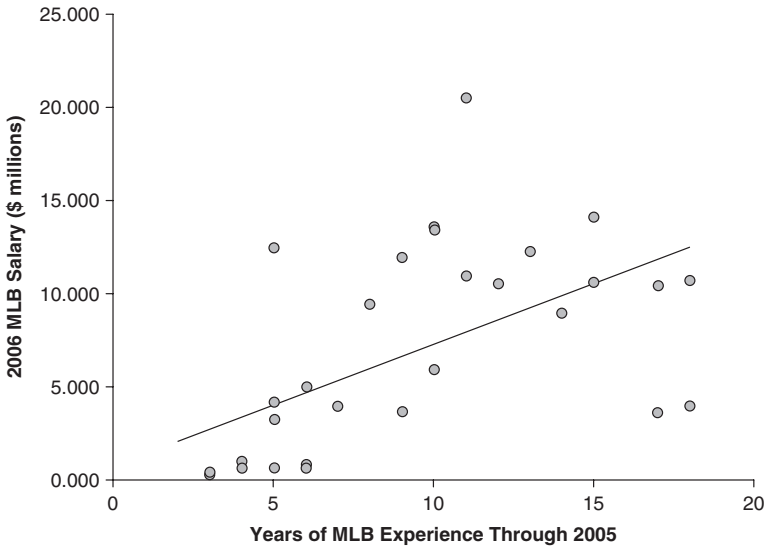


**Figure 2.3**

to do this calculation by hand in the previous example on baseball, these calculations can be performed more easily by using computer programs designed to carry out these tasks. There are *many* such programs that are capable of calculating sample regression lines.[6] Perhaps two of the most prevalent and commonly used programs are Microsoft Excel and SPSS. In order to give the reader experience in reading and interpreting regression analysis output from these programs, the examples presented from this point forward will use these two programs (on an alternating basis) to carry our calculations.[7]

Using Excel, we can input the data presented in Table A3 for the column labeled *VOTES* and the column labeled *GROWTH* into a spreadsheet. As shown at the bottom of Table A3, the variable *VOTES* is defined to be the percentage of the two-party vote received by the incumbent party candidate. This is our dependent variable $Y_t$ in Equation 1.6. The data for *GROWTH* are the growth

[6] For example, some popular programs include SAS, TSP, EVIEWS, MINITAB, SHAZAM, and STATA.

[7] Microsoft Excel is a relatively easy program to use and performs the basic calculations we will need for this book. SPSS is specifically designed to do statistical calculations and is capable of carrying out more sophisticated analyses. Appendix B provides some basic instruction on how to perform regression analysis using Excel and SPSS. The reader is referred to the instruction manual and tutorial that accompany these programs for greater details on how to use these programs. Additionally, Einspruch (2005) gives detailed instruction on how to use SPSS.

rates of real gross domestic product (GDP) over the three quarters prior to the election.[8] This is our independent variable $X_t$ shown in Equation 1.6.

Once the data have been input, we can then follow the steps for calculating the least-squares regression values for *a* and *b*. Doing so yields the results shown in Table 2.2.[9]

The first entry, "Observations," simply reports that we have 23 observations in our sample. Next, we see a column headed "Coefficients" and two rows labeled "Intercept" and "*GROWTH*." The entries are shown as 51.065 and 0.880 (rounding to three decimal places). These are the least-squares values for the intercept term, *a*, and the slope term, *b* (respectively), for Equation 1.6. Using these results, we can now write the predicted equation as

$$\hat{Y}_t = 51.065 + 0.880\ X_t. \tag{2.6a}$$

Once again, we may reexpress our results using variable names in place of *Y* and *X*:

$$\widehat{VOTES}_t = 51.065 + 0.880\ (GROWTH_t). \tag{2.6b}$$

This equation tells us that if the growth rate were zero over the three quarters prior to the election, then the incumbent party candidate is expected to receive approximately 51.065% of the two-party vote. In addition, for every 1 percentage point increase of real GDP over the three quarters prior to the election, the incumbent party candidate is expected to gain approximately 0.880 percentage points of the two-party vote. This, of course, works in the other direction as well. That is, for every 1 percentage point decline in the economy's real GDP, the incumbent party's candidate is expected to suffer a 0.880 percentage point loss of the two-party vote.

The last part of Table 2.2 shows predicted values and the associated residuals for 23 elections covered in our sample of data. The column headed "Predicted *VOTE*" shows the percentage of the two-party vote the incumbent party candidate was predicted to receive according to Equation 2.6b,

[8] The U.S. GDP is a measure used by economists to track the growth of the U.S. economy. It is defined as the total market value of all final goods and services produced inside the United States over a specified period of time.

[9] The results presented in Table 2.2 are an edited version of the actual output Excel produces. Much of the actual output the program produces was excluded in this example to make things easier for the reader. The full output will be presented in later examples as we progress.

**Table 2.2**

| SUMMARY OUTPUT | | |
|---|---|---|
| Observations | 23 | |
| | Coefficients | |
| Intercept | 51.065 | |
| *GROWTH* | 0.880 | |
| RESIDUAL OUTPUT | | |
| Observation | Predicted *VOTE* | Residuals |
| 1 | 53.000 | –1.300 |
| 2 | 40.947 | –4.847 |
| 3 | 47.634 | 10.566 |
| 4 | 55.112 | 3.688 |
| 5 | 38.220 | 2.580 |
| 6 | 61.358 | 1.142 |
| 7 | 54.232 | 0.768 |
| 8 | 54.936 | –1.136 |
| 9 | 53.616 | –1.216 |
| 10 | 51.769 | –7.169 |
| 11 | 49.833 | 7.967 |
| 12 | 51.417 | –1.517 |
| 13 | 55.552 | 5.748 |
| 14 | 55.552 | –5.952 |
| 15 | 56.432 | 5.368 |
| 16 | 54.584 | –5.684 |
| 17 | 47.898 | –3.198 |
| 18 | 55.992 | 3.208 |
| 19 | 53.088 | 0.812 |
| 20 | 53.000 | –6.500 |
| 21 | 53.440 | 1.260 |
| 22 | 52.472 | –2.172 |
| 23 | 53.616 | –2.416 |

plugging in the actual value of *GROWTH*. That is, for Observation 1 (the 1916 presidential election), the growth rate reported for that year is 2.2 (from Table A3). If we plug this value into Equation 2.6b, we would have the following predicted value for that election:

$$\widehat{VOTES}_1 = 51.065 + 0.880\ (2.2) = 53.000. \qquad (2.6c)$$

However, the actual value for that year was 51.7. The fact that our predicted value is not equal to the actual value again represents the fact that our very simple model is not capable of explaining the entire behavior of *VOTES*. The difference between actual values and the predicted values is what is shown in Table 2.2 as the "Residuals"—the "*e*" in our Equation 1.6. Thus, for the first observation, we have (rounding to one decimal place):

$$e_1 = VOTES_1 - \widehat{VOTES}_1 = 51.7 - 53.0 = -1.3. \qquad (2.7)$$

This result shows that our model overpredicted the actual percentage of the two-party vote received by the incumbent party candidate (Woodrow Wilson, in this case) by 1.3 percentage points. The residuals are calculated for each election in our sample and are reported in Table 2.2.

By plotting the regression line shown in Equation 2.6a, along with the actual values for $Y_t$ and $X_t$, we have Figure 2.4.[10] As can be seen, the actual

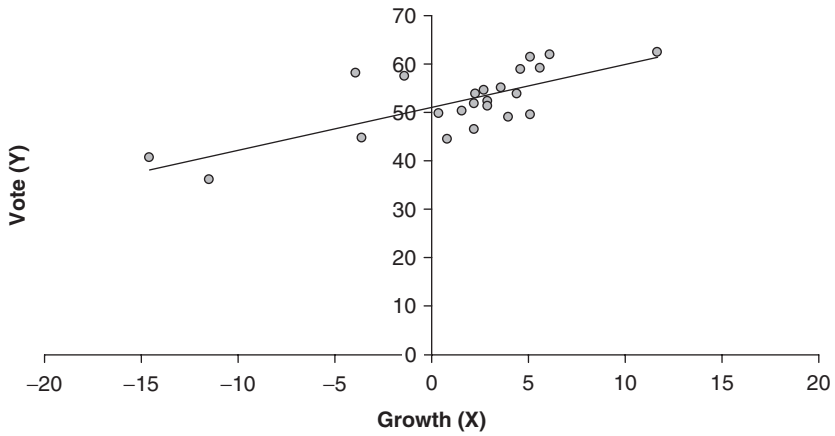


**Figure 2.4**

[10] This graph was created using the Excel program.

values do not fall precisely on the regression line, but are speckled above and below it. Again, this illustrates the imperfect relationship between $Y_t$ and $X_t$, with the vertical distance from any dot to the regression line equaling the error in prediction.

Turning now to our example of state abortion rates, we can use the same tool of least-squares to estimate the sample regression function shown in Equation 1.7. Recall that for this model, the dependent variable, $Y_i$, is the abortion rate for state $i$, and the independent variable, $X_i$, is the measure for the variable called *RELIGION*. In this case, we will use the program SPSS to calculate the least-squares values for $a$ and $b$ in Equation 1.7. In doing so, we obtain the output shown in Table 2.3.[11]

**Table 2.3**

**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

We see in Table 2.3, under the "Unstandardized Coefficients" heading, that the intercept, or "constant" term, is 23.825. This is our value for a in Equation 1.7. The value for $b$ is the coefficient for *RELIGION* and is shown to be –0.099. Using these values for $a$ and $b$, we have the following sample regression function (rounded to three decimal places):

$$\hat{Y}_i = 23.825 - 0.099\ X_i, \tag{2.8a}$$

or, using variable labels,

$$\widehat{ABORTION}_i = 23.825 - 0.099\ (RELIGION_i). \tag{2.8b}$$

This equation for the sample regression function is plotted in Figure 2.5[12] along with the actual data.

[11] As in the previous example using Excel, the actual output produced by SPSS was edited so as to include only the results relevant for our discussion at hand. The full SPSS output will be presented in later chapters.

[12] Figure 2.5 was produced by the SPSS program.

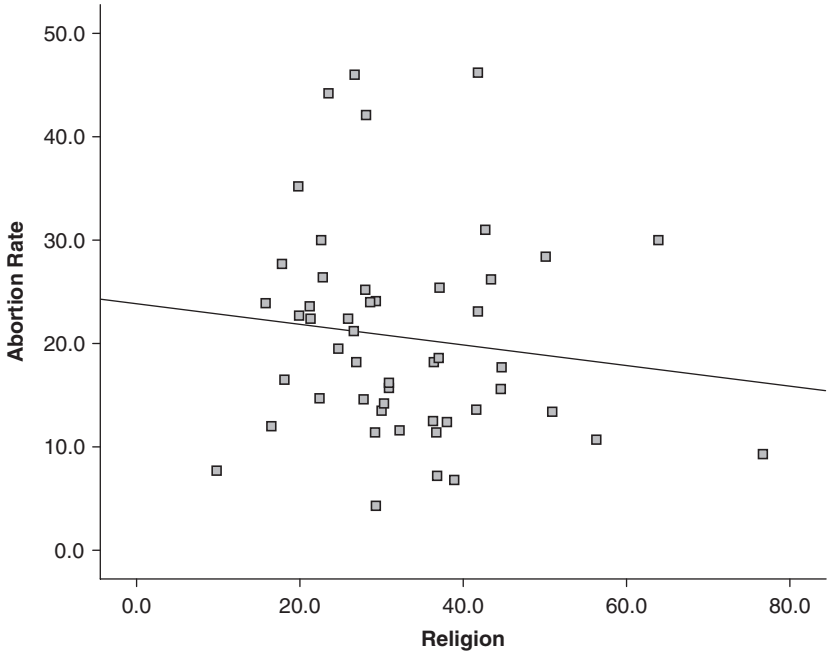


**Figure 2.5**

The squares representing actual observations are not closely speckled around the regression line, indicating that although $X_i$ (religion) may explain some of the behavior of $Y_i$ (the abortion rate), a great deal has been left unexplained. As we have discussed, this unexplained portion is captured by the error term $e_i$ and is equal to the vertical distance from the dots shown to the regression line.

This example illustrates the point made in Chapter 1 that in many cases, our simple bivariate model will not be sufficient for explaining the behavior of a dependent variable, and a multivariate regression model will be needed. The subject of multivariate models will be taken up later in Chapter 4.

We now move to our crime example. Using data for 42 British police force areas for 2004, we can estimate the model presented in Equation 1.8 using our OLS method. The results, produced by SPSS, are shown in Table 2.4.

Using the output from Table 2.4, we can write the estimated sample regression function as

$$\hat{Y}_i = 51.772 + 10.218X_i, \tag{2.9a}$$

**Table 2.4**

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | *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

or, alternatively,

$$\widehat{CRIME}_i = 51.772 + 10.218\ (UNEM_i). \tag{2.9b}$$

Recalling that *CRIME* is measured as the number of recorded crimes per 1,000 people, and *UNEM* is the male employment rate in percentages (both for the year 2004)[13], then interpretation of Equation 2.9b is straightforward. If *UNEM* were zero, then the estimated constant term tells us we would expect about 51.772 crimes per 1,000 people, all else being equal. The estimated coefficient to *UNEM* tells us that, all else being equal, a 1 percentage point increase in the male unemployment rate tends to increase crime by approximately 10.218 crimes per 1,000 people. In order to predict crime rates using Equation 2.9b, we can simply plug in a value for *UNEM* and compute the predicted value for *CRIME*. For example, we can consider the police force area "Avon & Somerset" (Avon for short) for 2004. The reported unemployment rate was 1.888%. Using this value in Equation 2.9b, we find

$$\widehat{CRIME}_{\text{Avon}} = 51.772 + 10.218\ (1.888) = 71.064. \tag{2.9b}$$

The actual reported crime rate for Avon in 2004, however, was 85.941 crimes per 1,000 people. Computing our error in prediction for 2004, we have

$$e_{\text{Avon}} = CRIME_{\text{Avon}} - \widehat{CRIME}_{\text{Avon}} = 85.941 - 71.064 = 14.877. \tag{2.10}$$

In this case, we observe a rather large, positive residual. This result illustrates the fact that our model for crime shown in Equation 1.8 is much too simplistic and that many other factors (not included in the model) are important in determining crime rates. Figure 2.6a graphs the sample regression

[13] The male unemployment rate is used here because the vast majority of all crimes are committed by males in their late teens to early twenties.

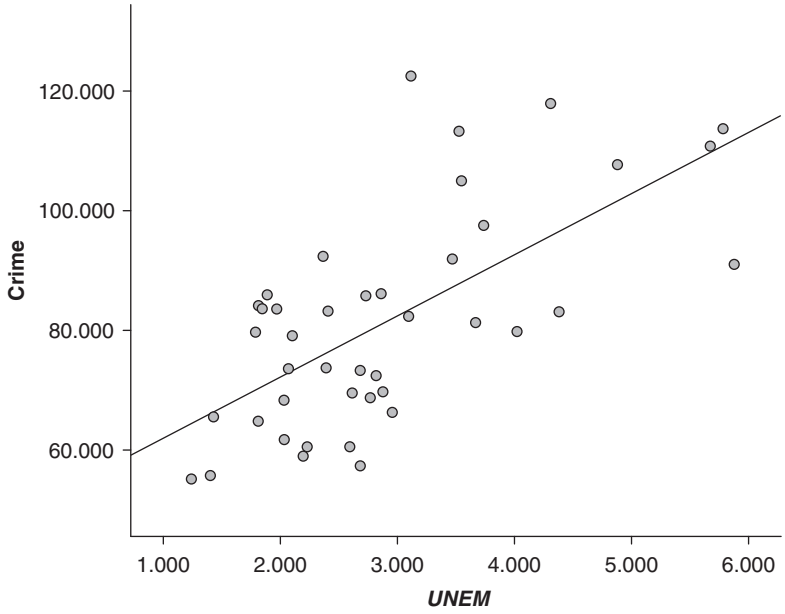


**Figure 2.6a**

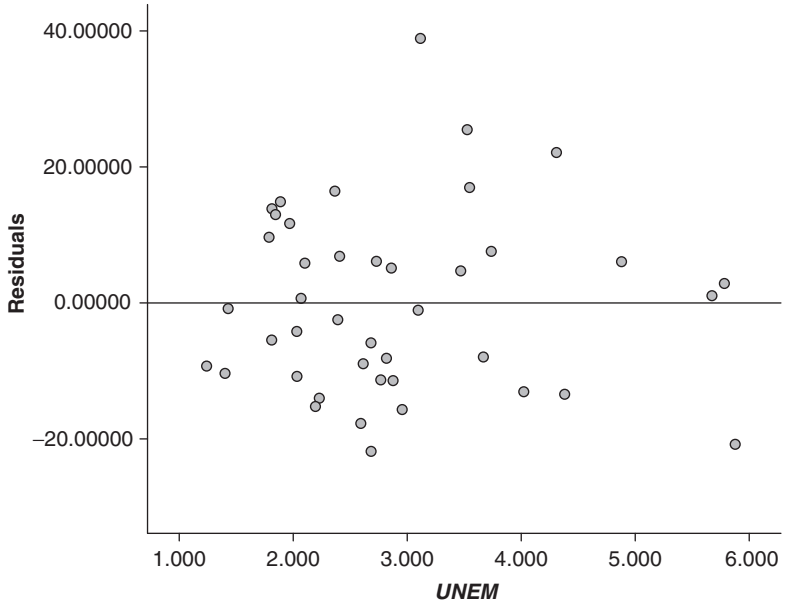


**Figure 2.6b**

function for Equation 2.9b and the actual values for 2004 crimes per 1,000 people and unemployment rates.

As one can see, the dots are widely spread around the sample regression function, further illustrating the limited ability of unemployment rates alone to explain the crime rates per 1,000 people across the 42 police force areas. Another visual method of showing the poor fit of the sample regression function is to plot the residuals against the unemployment rates, as shown in Figure 2.6b. A horizontal line corresponding with zero on the vertical axis is plotted along with the residuals. We can see that the residuals are fairly evenly spread above and below zero, but at the same time, the spread is quite wide. Again, this tells us that, although the unemployment rate may be able to partly explain crime rates, much is left unexplained.[14]

## Regression Model Assumptions and the Properties of OLS

The OLS method of estimating regression lines is clearly a powerful research tool. The validity of the OLS results we obtain, however, depends on a series of assumptions, called the Classical Linear Regression Model (CLRM) assumptions, which we have yet to discuss. These assumptions are sketched briefly below.[15] The end result is that if these assumptions are satisfied, then the OLS estimated regression line gives us the best possible representation of the population's regression line.[16]

### CLRM Assumptions

1. The average of the population errors ($u_i$s) is zero. As seen in Figure 1.2b, some points will lie above the population regression function and will have positive errors, and some will lie below and have negative errors. On average, the errors should cancel each other, and thus the average of the errors should be zero.

2. The spread of the errors above and below the regression line (i.e., the **variance**) is uniform for all values of $X$. Graphically, this means that the

[14] A more formal discussion of how well an estimated sample regression function explains the behavior of the dependent variable (such as crime rates) will be presented in the following chapter.

[15] For a more detailed discussion of the regression model assumptions, see Berry (1993) or Wooldridge (2006).

[16] A formal proof that the OLS method is the best one is given by the famous **Gauss-Markov theorem**. For more details, see Greene (2003).

actual observations for $Y_i$, for given values of $X_i$, fall within a uniform band around the population regression function, as seen in Figure 2.7. As shown, the population regression function (PRF) has observations that are above and below it, but they are uniformly spread around the line, as the two darker, parallel lines above and below the PRF demonstrate. (The technical term is that the errors are said to be **homoscedastic**, meaning "equal variance." An example of when the assumption is violated is given later in Chapter 7.)

3. The error associated with one observation is not associated with errors from any other observations. Or, in technical terms, we assume no **autocorrelation** among the error terms. The basis for this assumption is straightforward. The errors are supposed to represent purely random effects for which our model is unable to control. If, however, one observation's error is somehow related to another observation's error, then this implies that there is some systematic relationship among the errors, and thus they are not purely random. The implication is that this systematic relationship contains information which we should use to improve the estimation of our model. If it is ignored, then we are not fitting the best line to our data. (An example of this kind of problem is considered in Chapter 7.)

4. The independent variable $X_i$ is uncorrelated with the error term $u_i$. The reasoning behind this assumption can be understood if we recall our goal: to isolate the separate effects of changes in $X_i$ on $Y_i$. Suppose, for example, that $X_i$ and $u_i$ are positively correlated, meaning that an increase in $u_i$ would generally be associated with an increase in $X_i$. If this is the case, then it would be

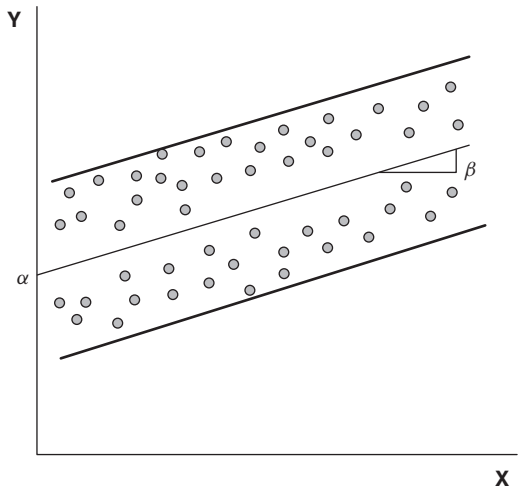


**Figure 2.7**

difficult to isolate the separate effects of an increase in $X_i$ on $Y_i$ because any increase in $X_i$ could be due, in part, to an increase in $u_i$. An example of when this assumption is violated is when we have an "omitted variable bias," a topic we touch upon in Chapter 7.

5. The variables $X_i$ and $Y_i$ contain no measurement errors. Again, the necessity of this assumption is easy to understand. If $X_i$ and $Y_i$ are measured inaccurately, then the OLS values for *a* and *b* (which, as Equations 2.3a and 2.3b show, are derived from the values of $X_i$ and $Y_{i,}$) are not likely to be accurate estimates of the population's $\alpha$ and $\beta$.

6. The model we put forth, such as the one in Equation 1.2b, is theoretically sound. This assumption can be violated in a number of ways, including omitting relevant independent variables. For example, if we consider our baseball salary model, we hypothesized that years of playing determine a player's salary. We know, however, that years alone do not determine baseball salaries. A player's offensive ability (e.g., hitting) or defensive ability (e.g., fielding) is also obviously important. If we do not take into account these important factors in our model, then our model will not be correctly specified (i.e., we commit a **model specification error**), and the values for a and b may not be reliable.

Another way we can misspecify our model is to use a functional form that is inappropriate. For example, we may fit a straight line to data when, in fact, a curve is appropriate. (We will have more to say about functional forms later in Chapter 5.)

7. Our last assumption has to do with how the population's error term, $u_i$, is distributed. We assume that the $u_i$ follows a normal distribution (often referred to as the **normality assumption**). That is, the random errors follow the familiar bell-shaped curve that is well known from statistics. The importance of this assumption will be seen later in Chapter 3.[17]

If CLRM assumptions 1 through 6 are satisfied, then, as noted above, the OLS regression line provides the best possible estimate of the population regression line. Or, using the more common terminology, OLS is **BLUE**, which is an acronym for Best Linear Unbiased Estimator.[18] In order to

[17] The justification of this assumption comes from a theory from statistics called the Central Limit Theorem. Those interested in learning more about this theorem are referred to Greene (2003).

[18] Assumption 7, the "normality assumption," is not required for the BLUE property of the OLS estimation method. It is made for purposes of hypothesis testing, a topic we take up in the next chapter.

understand this property, we can begin by discussing what is meant by **linear** and **unbiased**. Linear simply means that we are estimating a value for the intercept ($a$) and the slope term ($b$) that are raised only to the power 1. Thus, our Equation 1.5 is an example of a linear regression model. However, consider the following equation:

$$Y_i = a + b^2X_i + e_i. \tag{2.11}$$

In this case, $b$ is raised to the power 2 and thus is not "linear" in the sense described here. On the other hand, we do allow for the independent variables to enter into our equation nonlinearly, as shown in the following equation:

$$Y_i = a + bX_i^2 + e_i. \tag{2.12}$$

This is considered a linear estimation because $a$ and $b$ are raised to the power 1 and thus Equation 2.12 is a candidate for the BLUE property of OLS.[19]

Unbiasedness has to do with the fact that our sample estimate of $b$, the slope term, is a random variable. That is, given that we use a sample to calculate $b$, if we repeat the estimation with new samples, we will likely find different values for $b$. If we do so, we can then calculate *the average of all of these bs*. If it is true that the average of the $b$s is equal to the population's true $\beta$, then the estimator is said to be unbiased; similarly for $a$, the intercept term.[20]

Now that we know what is meant by linear and unbiased, we can explain what is meant by "best." If we consider all possible estimation methods that are linear and produce unbiased estimates of $a$ and $b$, the OLS method is the best one in the sense that it gives us the most precise estimates of $a$ and $b$. In order to understand the meaning of this statement, recall that the estimated parameters $a$ and $b$ "bounce" around from sample to sample (i.e., they are random variables). If they are unbiased, this means that as they bounce around, they have a mean value that is equal to the population's $\alpha$ and $\beta$. To be best, they will bounce around the least; any other linear unbiased estimation method will produce values of $a$ and $b$ that bounce around more than

[19] Our linearity requirement is only for the parameters to be estimated (i.e., $a$ and $b$). The use of nonlinear independent variables, as shown in Equation 2.12, is discussed later in Chapter 5.

[20] Formally, if it is true that $E(a) = \alpha$ and $E(b) = \beta$, where "$E$" stands for expected value, then we say that $a$ and $b$ are unbiased estimators of $\alpha$ and $\beta$. It then follows that if $E(b) \neq \beta$, then $b$ would be a biased estimator of $\beta$; similarly for the intercept term, $a$.

those calculated using the OLS method. In other words, of all linear unbiased estimation methods, the OLS method gives us the most precise estimates of $\alpha$ and $\beta$, or OLS is BLUE.[21]

Thus, we have powerful support for the use of OLS.

## Summing Up

In this chapter, we have seen how we may use samples of data and employ the method of least-squares to estimate the linear relationship between a dependent variable and an independent variable. We have seen, however, that our estimated sample regression function does not completely explain the relationship between the dependent variable and the independent variable, and what is left unexplained shows up in the error term. The question we now turn to is that of model performance and reliability. That is, once we have estimated a relationship between a dependent variable and an independent variable, what can we say about *how well* we have estimated this relationship? This is the topic of the next chapter.

## PROBLEMS

2.1 Consider the following data set:

| $Y_i$ | $X_i$ |
|---|---|
| 10.5 | 13 |
| 9.75 | 12 |
| 10.00 | 12 |
| 12.25 | 14 |
| 8.00 | 10 |

where: $Y_i$ is individual *i*'s hourly wage (in dollars per hour)
$X_i$ is individual *i*'s number of years of education

Use Equations 2.3a and 2.3b to calculate the OLS values for *a* and *b* and interpret your results.

2.2 Use Equation 2.5b and the data in Table A1 of Appendix A to predict the salary of Javy Lopez. What is the error in prediction (i.e., $e_i$)? What may account for this error?

[21] In technical terms, OLS estimates of the random variables *a* and *b* will have the smallest variance as compared to any other estimation method. For a more detailed discussion, see Gujarati (2003) or Wooldridge (2006).

2.3 Suppose we have the following model:

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

where, in this case, $Y_i$ is the manufacturer's suggested retail price (*MSRP*) for a sports utility vehicle (SUV) and $X_i$ is the horsepower of the SUV.

a. What sign do you expect for $\alpha$ and $\beta$?
b. Using SPSS or Excel (or an equivalent program), and the data provided in Table A5 in Appendix A, perform an OLS regression for the above model and interpret the estimated coefficients.

2.4 Recall the model discussed in Problem 1.3, which shows students' performance in college as a function of their SAT scores:

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

where: $Y_i$ is individual *i*'s freshman college grade point average (GPA)
$X_i$ is individual *i*'s SAT score

a. Using Excel or SPSS (or an equivalent program) and the "GPA" data set, create a plot of the data with the variable GPA on the *Y*-axis and SAT on the *X*-axis. Does there appear to be any relationship between the two?
b. Using SPSS or Excel (or an equivalent program), perform an OLS regression with GPA as the dependent variable and SAT as the independent variable. Interpret the estimated constant term and coefficient to SAT.