1

An Introduction to the Linear Regression Model

The basic goal of regression analysis is to use data to analyze relationships. Thus, the starting point for any regression analysis is to have something to analyze. That is, we begin with some idea or hypothesis we want to test and we then gather data and analyze these data to see if our idea is verified. The purpose of this chapter is to provide the reader with several examples of the kind of research that can be done with regression analysis techniques. These examples, which are woven throughout this book, were chosen in such a way as to illustrate to the reader how regression analysis methods can be used to understand relationships across a broad range of subjects. Once we understand the basic notion of regression analysis, we then proceed to Chapter 2, where the more technical aspects of regression analysis are discussed.

Baseball Salaries

Suppose we are interested in exploring the factors that determine one's salary. There are many such factors, one of which would be the experience an individual has in his or her profession. That is, for most professions, the longer a person has been on the job, the greater is his or her salary. The logic behind this relationship is that workers learn with experience and become more productive over time. As such, employers reward workers for their increased productivity that comes with experience. But how large is the reward

in relation to increased experience? That is, as a person gains another year of work experience, by how much can he or she expect his or her salary to increase from one year to the next? One method of trying to understand this relationship between salary and experience is to collect data on individuals within a profession and use a graph to visualize the relationship between the two. As an example, let's consider the occupation of professional baseball players.

Salaries earned by Major League Baseball (MLB) players have been the subject of great discussion in the media largely because in recent years, players have earned enormous amounts of money for playing the game. We may consider, then, how a player's salary is related to his experience in MLB.[1] Data on players' salaries have become public information these days, as a number of media sources publish the earnings of players as well as other information about them, such as years of MLB experience.[2] Suppose we collected a sample of data on player salary and experience and plotted these

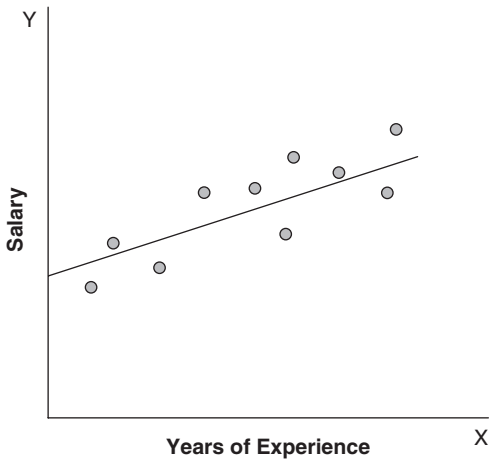

Figure 1.1

[1] There has been, in fact, a great deal of empirical research done on this topic; see, for example, Scully (1974) and Zimbalist (1992).

[2] For example, data on player salaries are published annually by a number of newspapers (e.g., *USA Today*) and are also available online at various Internet sites, including Sean Lahman's "Baseball Archive," which is located at www.baseball1.com. Information on player experience and performance can also be found at this Web site and is published annually in various other sources, including Thorn and Palmer's *Total Baseball* (1997).

pairs of numbers on a graph with a player's salary on the vertical, or *Y*, axis and the corresponding years of experience on the horizontal, or *X*, axis. Figure 1.1 shows an example of how this graph may look.

Viewing Figure 1.1, we can observe that the collection of dots, each of which represents an individual player's salary and his associated experience, tend to rise as we move out along the *X* axis. As a means of trying to represent the general behavior of these dots, a line has been run through them that shows their general tendency to rise. As the line suggests, as a player's experience (*X*) increases, his pay (*Y*) tends to increase as well. This would seem to support our hypothesis that workers (players) are rewarded with greater salaries as their experience (years of playing in MLB) increases.

By adding a line to our Figure 1.1, we were able to capture the general relationship between salary and experience. But in doing so, it also implies a more specific assumption about the behavior of *Y* with respect to *X*. This assumption, known as the linear regression model assumption, forms the basis for regression analysis and is explained below.

Linear Regression Model Assumption

The easiest way to understand the linear regression model assumption is to illustrate it with an example. Returning to our case of baseball, suppose instead of just a sample of data, we collect data for *all* MLB players. Having such a large collection of data, we could then order our data such that players are grouped according to the number of years of MLB experience, which was our *X* variable for this example. Thus, all players with, say, 1 year of experience would be grouped together. All players with 2 years of experience would be in another group, and so on. We could record the salary of each player in each group, and then use this information to calculate the average salary for each group as well. This procedure is illustrated graphically in Figure 1.2a.

Viewing Figure 1.2a, we can consider players with 1 year of MLB experience who have their salary plotted on the graph above the value shown as 1 on the *X* axis. Notice that some players in this group have higher salaries than others, perhaps because of differences in other skills (this point is expanded on later). If we calculated the average salary of players in this group, its value would lie somewhere in the middle of these plotted points, such as the point shown with a heavier dot. Thus, this heavy dot represents the mean or average salary of players with 1 year of experience. We can carry out this same exercise for players with 2 years of MLB experience. These individuals have their salary plotted above the value of 2 on the *X* axis. As in the previous case, some players in this group will have higher salaries than others, and the average

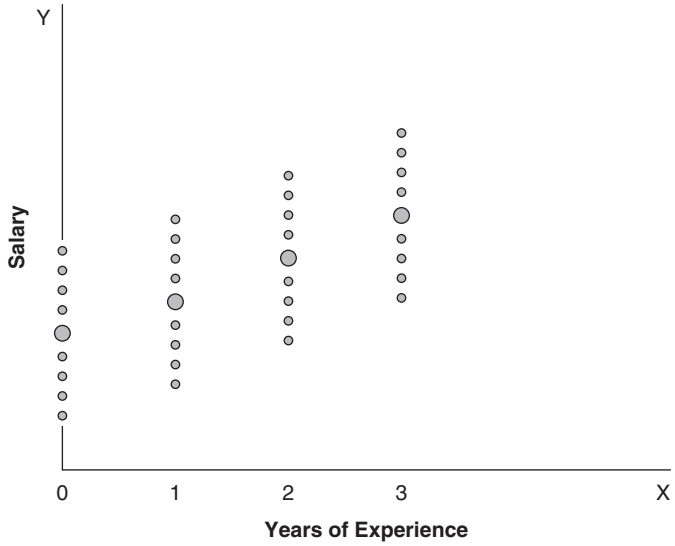

Figure 1.2a

salary for all players with 2 years of MLB experience is shown by the heavy dot above 2 on the *X* axis. This same kind of analysis can be done for players with 3 years of MLB experience, and the heavy dot above the value of 3 on the *X* axis represents the mean salary for all players in this group.

This procedure can, in fact, be done for all values of **X**, MLB experience, in each case calculating the mean value for salary (*Y*) for given values of experience (*X*). Given this graph, we have the following assumption: *The linear regression model assumes that the mean values of Y, for given values of X, are a linear function of X*. Or, in terms of our graph, the heavy dots (which are the mean values of *Y* for given values of *X*) lie on a line. (It should be noted that in some cases, the relationship between the mean values of ***Y*** and *X* may be *non*linear. Examples of nonlinear relationships are discussed in Chapter 5.) This assumption is shown graphically in Figure 1.2b, which takes Figure 1.2a and adds a line connecting the heavy dots.

This assumption can also be expressed somewhat more formally by using the following mathematical expression:

$$E(Y \mid X_i) = \alpha + \beta X_i. \tag{1.1}$$

The *E* in Equation 1.1 stands for "expected value" or mean, and the vertical line, $\mid$, can be read as "for given values" of X_i. (The subscript *i* is used

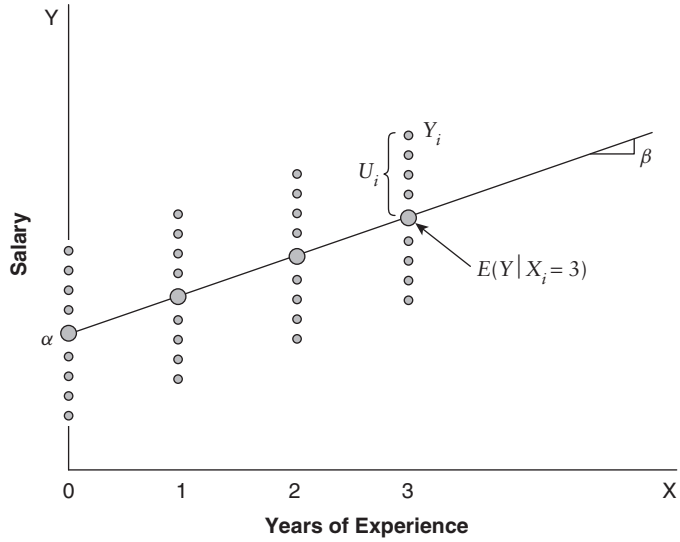

Figure 1.2b

to keep track of different values that X can take on.) The expression on the right of the equal sign, $\alpha + \beta X_i$, is simply the equation to a line. Or, putting it all together, Equation 1.1 can be read as: "the expected value of Y for given values of X_i is equal to a linear function of X_i." As for terminology, the variable Y is called the **dependent variable** because its value is said to depend on the value that X_i, which is called the **independent variable**, takes on.[3] The symbols α and β in Equation 1.1 are constants and are referred to as the intercept and slope terms, respectively (it is common in regression analysis to use Greek letters such as these). The intercept term α tells us what the expected value of Y (in our case, salary) would be for individuals who have no experience (i.e., new, or rookie, MLB players). That is, if a player has no experience, then his value for X_i is zero. Plugging in zero for X_i in Equation 1.1, we have

$$E(Y \mid X_i = 0) = \alpha + \beta(0) = \alpha. \tag{1.1a}$$

This is shown in Figure 1.2(b), where a line representing Equation 1.1 intersects the vertical axis.

[3] The independent variables are also sometimes referred to as "predictors" or "explanatory variables."

The slope term β in Equation 1.1 tells us how *Y* is expected to change for each one-unit increase in X_i. Or, in the case of MLB salaries, how salaries are expected to change for each additional year of experience. To see this, consider a player with 1 year of MLB experience. His value of X_i then is 1, and plugging this into Equation 1.1 yields

$$E(Y \mid X_i = 1) = \alpha + \beta(1) = \alpha + \beta. \tag{1.1b}$$

Comparing Equations 1.1a and 1.1b, we see that the difference between the two is that players with 1 year of experience are expected to have β more in salary than players with no experience. In the case of players with 2 years of experience, they are expected to have 2β more in salary as compared to those with no experience. Thus each additional year's worth of experience increases a player's expected salary by β.

The implication of writing the equation with $E(Y \mid X_i)$ is that it implies the understanding that *individual* values for *Y*, for given values of X_i, will not likely be exactly equal to $\alpha + \beta X_i$. To see this, we can return to Figure 1.2(b) and consider player i, who has 3 years of experience in MLB and earns a salary of Y_i. Notice that for this individual player, his actual salary Y_i is greater than the mean salary for his group, shown as $E(Y \mid X_i = 3)$. The difference between the actual and expected value for *Y* is shown as u_i. In terms of an equation, we can write a player's actual salary as

$$Y_i = E(Y \mid X_i) + u_i. \tag{1.2a}$$

Or, using Equation 1.1, we can rewrite the last equation by replacing $E(Y \mid X_i)$, giving us

$$Y_i = \alpha + \beta X_i + u_i. \tag{1.2b}$$

Every dot shown in Figure 1.2b, which represents a particular player's experience and his actual salary, can be expressed in a similar way. That is, every individual player's salary can be expressed as the sum of his group's expected salary, plus the specific player's value for u_i. What does u_i represent? The term u_i, which is called the **error term**, represents all the other factors that may affect player *i*'s salary that are not taken into account by the simple model shown in Equation 1.1. There are, in fact, numerous other factors that enter into the determination of salaries. In baseball, for example, players are rewarded for their offensive (e.g., hitting) and defensive (e.g., fielding) abilities. The fact that these other important explanatory variables are not accounted for in our model means that player salaries would not

likely fall exactly on the line shown in Figure 1.2b. To further illustrate why this is the case, consider two players who are identical in all measures, including years of experience, except that one player is a better hitter. This being the case, the better hitter would likely earn a greater salary because he is worth more to a team. What this means, then, is that although a player's experience may be an important factor in explaining his salary, experience alone cannot perfectly explain a player's salary. The error term included in Equation 1.2b is said to be **stochastic**, meaning that it is a random component of a player's salary, which varies from one player to another. Thus, if we again consider our specific player *i*, who has $X_i = 3$ years of playing experience, we see in Figure 1.2b that the vertical distance from the heavy dot on the line to the point representing this player's salary is the positive error u_i. This means that our player *i* is paid more than expected, perhaps because he is a better hitter, a factor not taken into account in our simple model. In a similar way, points below the line represent players whose salaries are less than expected (i.e., they have negative errors), perhaps because they are below-average hitters.

At this point, the reader may be wondering if it is possible to build a more elaborate model that takes into account these other factors that are missing from our model and that end up in the error term u_i. To a certain extent, this can and will be done in later chapters when we build on this simple model to include other explanatory measures such as hitting and fielding. In any case, it is not likely that *all* factors can be accounted for so that the error term is driven to zero.[4] This is true for a number of reasons. First of all, there may not be data available for many important variables (e.g., a player's speed in running the bases). Second, some factors that affect a player's salary may not be measurable (e.g., leadership ability or fan appeal). All of these factors that are not accounted for in our model end up in the error term, which will vary from player to player.

For now, we will continue to work with simple models like that shown in Equation 1.2b, which are referred to as **two-variable linear regression models** (also known as **bivariate linear regression models**) because they

[4] There is a case when the error term will, in fact, be zero. This is when an identity has been estimated. For example, suppose we collect data for distance measurements in meters and then collect data for the same distance measurements in inches. If we tried to estimate the relationship between meters and inches, we would find a perfect linear relationship and the errors would all be zero. This is the result because 1 meter is defined to be exactly 39.37 inches, and if measurements are made carefully enough, there should be no errors. There is no reason, however, to estimate an identity because these relationships are already known.

include only an intercept (α) and one slope term (β). These simple models will serve as a starting point from which we can discuss many of the issues regarding regression analysis. Bear in mind, though, that in most cases, a two-variable model will be too simplistic for our purposes and a more complex model will be needed.

Population Data Versus Sample Data

Before moving on, we need to clarify some aspects of our data sets, namely, their size. Typically, when we consider a theory, such as MLB salaries as a function of years of experience, there is a relevant population of data. In the baseball example, it may be all MLB players, past and present. For this population of data, when we formulate a mathematical model for the behavior of a dependent variable as a function of an independent variable, we are constructing what is called the **population regression function** (PRF) because it presents a hypothesis about the behavior of the population of data. Thus, for MLB, the model shown in Equation 1.1 is a population regression function for salary determination in MLB. In most cases, however, it is not possible to collect data for the entire population, perhaps because the data do not exist or because it would be practically impossible to collect the data.[5] As such, samples of data are collected from the population and analyzed with the hope that the information contained in the sample is a good representation of how the population behaves. In order to keep the distinction between sample analysis and population analysis clear, we will use the following **sample regression function** (SRF):

$$\hat{Y}_i = a + bX_i, \tag{1.3}$$

where $\hat{Y}_i$ is the sample version of the expression $E(Y \mid X_i)$, and a and b are the sample versions of the population's α and β. Figure 1.3 shows a graph of the sample regression function. As in the case for the population, given our sample, we can represent a specific player i's salary as the sum of what our model predicts his salary to be based on his experience, plus the error in prediction, e_i:

$$Y_i = \hat{Y}_i + e_i. \tag{1.4}$$

[5] Suppose, for example, we were studying the eating habits of the U.S. population. It would be nearly impossible to collect information from every individual given that the U.S. population is approximately 300 million.

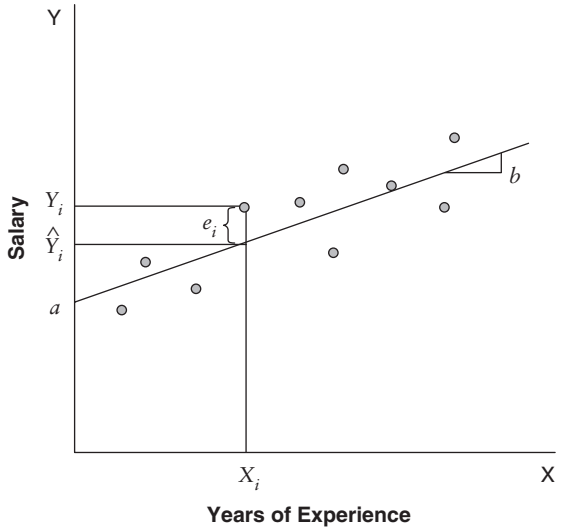

Figure 1.3

Using Equation 1.3, we can rewrite this expression by substituting for $\hat{Y}_i$, giving us

$$Y_i = a + bX_i + e_i. \tag{1.5}$$

Thus, Equation 1.5 shows the linear relationship between Y_i and $\mathbf{X}_i$, with the term e_i representing all other factors not accounted for in our model. This equation will be used in place of Equation 1.2b, which was for the population data, and the intercept term a is a sample estimate of the population's α and the slope term b is a sample estimate of the population's β. The term e_i is the error term (also called the **residual**) for our sample regression function and is analogous to the population's error term u_i.[6] As can be seen in Figure 1.3, the residual is simply the difference between player i's actual salary, Y_i, and the salary we would predict for a player with $\mathbf{X}_i$ years of experience, $\hat{Y}_i$ (i.e., the point on the line above $\mathbf{X}_i$).

Hopefully, the sample's intercept and slope terms closely resemble the population's parameters α and β. If this is the case, then we can be confident

[6] Some authors reserve the term "residual" only for the sample regression function's error term (e_i) and use "error" or "disturbance" for the population regression function's error term (u_i). We will use both residual and error terms for e_i, remembering that these refer to sample results.

that by analyzing the sample's values for these parameters, we can understand the behavior of the population.

Presidential Elections

As a second example of a regression analysis model, we can consider the topic of presidential elections. Some academics, such as Yale economist Ray C. Fair, argue that the state of the economy is an important factor in describing the voting pattern in presidential elections (Fair, 1996; see also Kramer, 1971; Stigler, 1973). As Fair (1996) puts it, "Voters hold the party in the White House responsible for the state of the economy" (p. 90).[7] For example, if the current president is a Democrat, and the economy has grown substantially over his term, then the party in power is given partial credit for that economic success and voters would then reward the Democratic presidential candidate with votes. On the other hand, if the economy has suffered from recession in the years prior to the election, the reverse is true and the incumbent party candidate suffers. This theory can be evaluated using regression analysis. We can model voting for incumbent party candidates with the following sample regression function:

$$Y_t = a + bX_t + e_t. \tag{1.6}$$

In Equation 1.6, we now have the dependent variable, Y_t, representing the percentage of the two-party votes received by the candidate running for president who belongs to the same party as the incumbent (note that this could be the incumbent himself if he is running for a second term, such as Ronald Reagan, who ran for reelection in 1984, and Bill Clinton, who ran for reelection in 1996). The variable $\boldsymbol{X}_t$ now represents the economy's real percent growth rate over some specified period prior to the election at hand.[8] In this case, the error term, e_t, represents other factors not taken into account, such as the inflation rate prior to the election and perhaps other, immeasurable factors such as charisma of the candidate. Finally, note that in this case, we use the subscript t (as opposed to i used for the baseball example) to distinguish individual cases because now we are considering results of elections at different points in time. Graphically, this model would look similar to the one

[7] Indeed, Bill Clinton's 1992 presidential campaign used the phrase, "It's the economy, stupid!"

[8] The "real" growth rate is a term economists use to refer to the economy's growth rate adjusted for inflation.

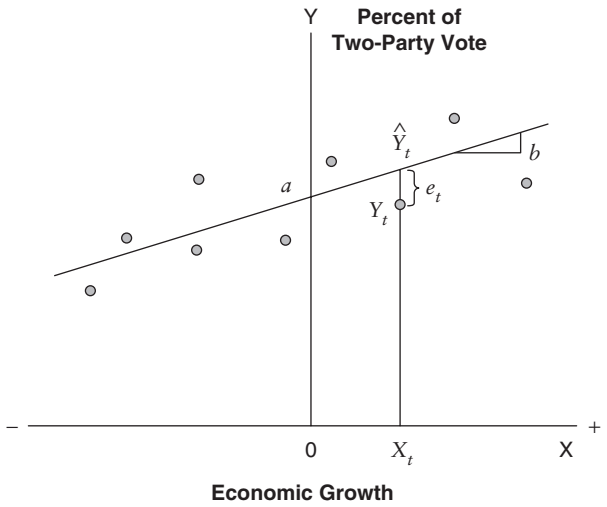

Figure 1.4

shown in Figure 1.3, except that the values for X_t, the real growth rate of the economy, can be negative. This is shown in Figure 1.4, which has a positive and negative range for X_t.

The value of a in this case would be the expected share of the two-party presidential vote received by the incumbent party candidate when the economy experienced no real growth (i.e., when $X_t = 0$). As for b, this would represent the increase (decrease) in the share of votes the incumbent party candidate would receive for a 1 percentage point increase (decrease) in the real growth rate. As in the case of baseball salaries, the actual data for Y_t would not likely fall exactly on the line, but would be "speckled" above and below the line as is shown in Figure 1.4. The vertical distance from these observations to the line shown would be the error term e_t, which, again, represents other factors affecting the share of the two-party votes the incumbent candidate received that are not taken into account in our simple model. As seen in Figure 1.4, for the given value X_t, the actual value of Y_t lies below the value predicted by the regression line, $\hat{Y}_t$. Thus, the associated error term e_t, which is equal to the actual value of Y_t minus its predicted value, would be negative.

Abortion Rates

Our third example of a regression analysis model deals with the socially sensitive issue of abortion. Abortion rates (the number of abortions performed

per 1,000 women of childbearing age) differ, sometimes greatly, across the United States. Researchers have been interested in discovering what factors play a role in explaining why, in some states, the abortion rate may be relatively high, whereas in others it is relatively low. There are, of course, many factors that affect the abortion rate across states, but one of them would likely be the moral views of the state's residents. Other things being equal, the greater the moral aversion to abortion, the fewer would be expected to be performed.[9] The moral position that residents of a state hold with regard to abortion is difficult to measure. One way to measure it is to consider what percentage of the state's population belongs to the Catholic, Southern Baptist, Evangelical, or Mormon faiths. These are the four main religions that have a stated opposition to abortion. Using this measure, which we will call "religion," we would expect that if we compare states, those with a greater percentage of state population that belongs to one of these faiths would tend to have fewer abortions, other things being equal. This relationship can be expressed, again, using an equation like we have seen in our previous examples. In this case, Y_i would be the abortion rate in state i, and $\boldsymbol{X}_i$ would be the measure for "religion," equal to the percentage of a state's population that belongs to one of the four faiths mentioned above. In this case, the slope term, β, would be negative, indicating that states with a relatively large value for "religion" would tend to have a lower abortion rate, all else being equal. That is, we would have the following sample regression equation:

$$Y_i = a + bX_i + e_i. \tag{1.7}$$

In this case, the error term e_i would capture other factors omitted, such as income and legal differences across the 50 U.S. states. It should be noted that in this example, the subscript i is used to keep track of values for $\boldsymbol{Y}$ and $\boldsymbol{X}$ for states (not individuals, as was the case in the baseball example). Graphically, we would have something like Figure 1.5. As in the other examples, the vertical distance from a given point on the graph (e.g., Y_i, $\boldsymbol{X}_i$) to the line would represent the error term e_i. In the example shown in Figure 1.5, the actual observation for the dependent variable, Y_i, lies below the predicted one, $\hat{Y}_i$, for the given value of the independent variable, $\boldsymbol{X}_i$, and so the error term would be negative.

[9] Previous research on the determinants of abortion rates can be found in Medoff (1988) and Kahane (2000).

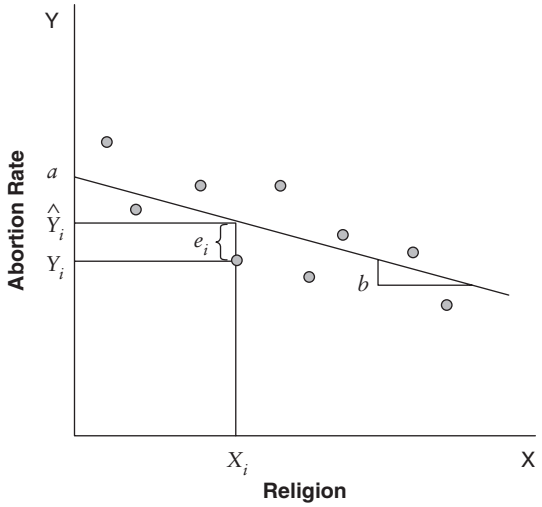

Figure 1.5

As for the intercept, *a*, it requires some additional discussion in this example. Technically speaking, the intercept represents the abortion rate that we would expect in the case where the variable religion is zero (i.e., $\boldsymbol{X}_i = 0$). However, intuition would tell us that it is clearly not the case that the variable religion would take on the value of zero in any state, as this would require that *none* of a state's residents belonged to the Catholic, Southern Baptist, Evangelical, or Mormon faiths. As we can see in Appendix A, which presents the data for the variable religion, our intuition is correct in that the value of religion is not, in fact, zero in any state. What this means for our model shown in Equation 1.7 is that for this example, the intercept has no sensible interpretation. That is, the intercept term is technically necessary to "anchor" the line in the graph, but beyond that it is meaningless. (This fact that the intercept may have no meaningful interpretation is often the case in regression model analyses.)

As for the value for the slope term, *b*, in this case it represents the predicted change in the abortion rate as the measure for religion increases by 1 percentage point.

Crime Rates

Our next example deals with the issue of crime. Much of the modern theory on the determinants of crime can be traced back to the seminal 1968 work

by economist and Nobel Laureate Gary Becker.[10] Becker viewed crime as a rational choice that individuals make after considering the costs and benefits of legal work versus criminal activities. As part of the computation of the benefits of legal work, individuals must consider the chances that legal work can indeed be obtained. A measure that can be used to gauge the ability that legal work may be available is the unemployment rate. Other things being the same, the higher the unemployment rate, the lower the chances are that legal work is available to the individual and hence, the greater the likelihood that the individual would pursue criminal activities. In terms of a sample regression equation, we would have

$$Y_i = a + bX_i + e_i, \tag{1.8}$$

where Y_i would be a measure of the crime rate (e.g., total crimes per 1,000 people) in location i and $\boldsymbol{X}_i$ would be the unemployment rate (in percent) in location i. In this case, the intercept term, a, would be the expected crime rate in the case where the unemployment rate, $\boldsymbol{X}_i$, is zero. The slope term, b, would then represent the predicted change in the crime rate (e.g., change in the number of crimes per 1,000 people) for a 1 percentage point increase in the unemployment rate (see Figure 1.6).

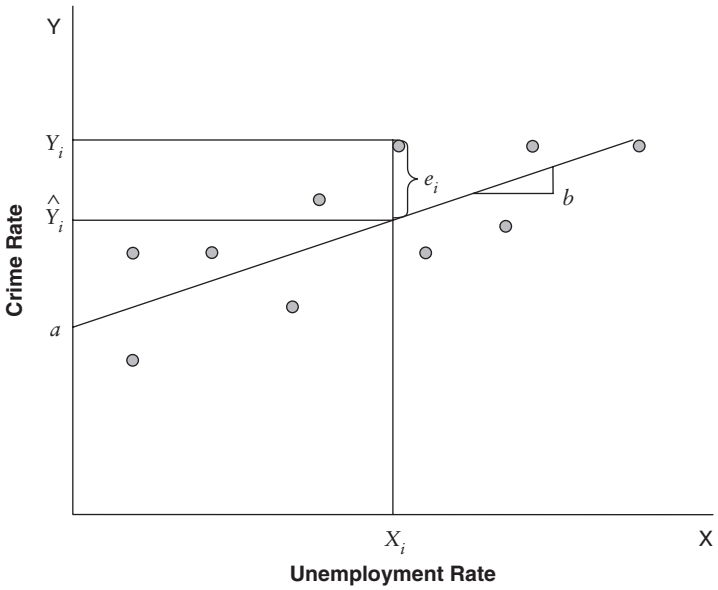

Figure 1.6

[10] See Freeman (1999) for a survey of research on the economics of crime.

In this case, the error term, e_i, represents other factors that may affect crime rates (e.g., police presence) that are not taken into account by the model.

Thus, we have four interesting examples that we can use to develop the basics of linear regression analysis. The graphs presented above for our four models show a line drawn through a scatter of points. At this point, two questions arise. First, what is the ultimate use of knowing such a line? The answer to this question is that, if we know this line, it can be used to predict values of the dependent variable, for given values of independent variables. For example, considering our model of MLB baseball salaries, if we know the values (i.e., the numbers) of the intercept, a, and the slope term, b, we could then use the line defined by them to predict a player's salary given his experience. This could be useful, for example, if a player is interested in knowing what his salary is expected to be as his experience increases. We will see in the following chapter how this kind of prediction can be accomplished. Of course, we must keep in mind that the predictions we make may not be entirely accurate because other factors that may be important in explaining a player's salary are not being considered. As we have already discussed, these excluded factors end up in the error term, e_i.

The second question that arises is, how can we find values for a and b so that the defined line "best fits" the scatter of points, which are our actual data? Recall that in our three examples of sample regression functions, we simply added a line to our graphs in such a way that the line seemed to fit the data well. This visual method, however, is imprecise, and there are better methods for accomplishing this task. This question of how we find the best fitting line to the data is, in fact, the subject of the next chapter.

Types of Data Sets

Finally, before moving on to the next chapter, we must say a few things about the various types of data sets. There are essentially three general varieties: **cross-sectional**, **time series**, and **pooled**. A cross-sectional data set fixes a point in time and looks across space. Our baseball example is a cross section because we collect data on salary and experience for a particular year and consider how salaries differ across players. In addition, our abortion example is a cross section because we consider a single point in time and look across the 50 states. A time series follows variables across time, while holding space constant. Thus, our presidential election example is a time series because we follow the breakdown of votes from one election to another. A pooled data set is a combination of both. For example, if we followed baseball salaries paid to all players *and* from year to year, we would have a

pooled data set. In this case, if we follow the *same* set of players from year to year, then this represents what is called a **panel data set.**[11] Panel data sets are very rich data sets, but they often require special treatment. We consider some simple methods of how to work with panel sets in Chapter 6.[12]

PROBLEMS

1.1 Consider the following model:

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

where: Y_i is individual i's wage

X_i is individual i's years of education

a. What is the interpretation of α? Do you expect it to be positive or negative?
b. What is the interpretation of β? Do you expect it to be positive or negative?
c. What does the error term, u_i, capture in this case?

1.2 Consider the model for presidential elections shown in Equation 1.6, which shows the percentage of two-party votes received by the incumbent party candidate. What other factors might be important in determining Y_t besides the real growth rate?

1.3 There has been considerable research into the relevance of SAT scores as predictors of students' performance in college (e.g., see the research by Bridgeman, McCamley, & Ervin, 2000; Camara & Echternacht, 2000; and Rothstein, 2004). We can consider this issue with the following model:

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

where: Y_i is individual i's freshman college grade point average (GPA)

X_i is individual i's SAT score

a. What is the interpretation of β? Do you expect it to be positive or negative?
b. What does the error term, u_i, capture in this case?

[11] If we took a different sample of players each year for a number of years, this would be termed a "pooled cross-sectional" data set.

[12] Greene (2003) and Wooldridge (2002) provide advanced discussions on the handling of pooled data sets.