CHAPTER 1

An Overview of Regression Analysis

1.1 What Is Econometrics?

"Econometrics is too mathematical; it's the reason my best friend isn't majoring in economics."

"There are two things you don't want to see in the making—sausage and econometric research."[1]

"Econometrics may be defined as the quantitative analysis of actual economic phenomena."[2]

"It's my experience that 'economy-tricks' is usually nothing more than a justification of what the author believed before the research was begun."

Obviously, econometrics means different things to different people. To beginning students, it may seem as if econometrics is an overly complex obstacle to an otherwise useful education. To skeptical observers, econometric results should be trusted only when the steps that produced those results are completely known. To professionals in the field, econometrics is a fascinating set

1. Attributed to Edward E. Leamer.

2. Paul A. Samuelson, T. C. Koopmans, and J. R. Stone, "Report of the Evaluative Committee for *Econometrica*," *Econometrica*, 1954, p. 141.

of techniques that allows the measurement and analysis of economic phenomena and the prediction of future economic trends.

You're probably thinking that such diverse points of view sound like the statements of blind people trying to describe an elephant based on what they happen to be touching, and you're partially right. Econometrics has both a formal definition and a larger context. Although you can easily memorize the formal definition, you'll get the complete picture only by understanding the many uses of and alternative approaches to econometrics.

That said, we need a formal definition. **Econometrics,** literally "economic measurement," is the quantitative measurement and analysis of actual economic and business phenomena. It attempts to quantify economic reality and bridge the gap between the abstract world of economic theory and the real world of human activity. To many students, these worlds may seem far apart. On the one hand, economists theorize equilibrium prices based on carefully conceived marginal costs and marginal revenues; on the other, many firms seem to operate as though they have never heard of such concepts. Econometrics allows us to examine data and to quantify the actions of firms, consumers, and governments. Such measurements have a number of different uses, and an examination of these uses is the first step to understanding econometrics.

1.1.1 Uses of Econometrics

Econometrics has three major uses:

1. describing economic reality
2. testing hypotheses about economic theory
3. forecasting future economic activity

The simplest use of econometrics is **description.** We can use econometrics to quantify economic activity because econometrics allows us to put numbers in equations that previously contained only abstract symbols. For example, consumer demand for a particular commodity often can be thought of as a relationship between the quantity demanded (Q) and the commodity's price (P), the price of a substitute good (P_s), and disposable income (Yd). For most goods, the relationship between consumption and disposable income is expected to be positive, because an increase in disposable income will be associated with an increase in the consumption of the good. Econometrics actually allows us to estimate that relationship based upon past consumption, income, and prices. In other words, a general and purely theoretical

$$Q = f(P, P_s, Yd) \tag{1.1}$$

can become explicit:

$$Q = 31.50 - 0.73P + 0.11P_s + 0.23Yd \tag{1.2}$$

This technique gives a much more specific and descriptive picture of the function.[3] Let's compare Equations 1.1 and 1.2. Instead of expecting consumption merely to "increase" if there is an increase in disposable income, Equation 1.2 allows us to expect an increase of a specific amount (0.23 units for each unit of increased disposable income). The number 0.23 is called an estimated regression coefficient, and it is the ability to estimate these coefficients that makes econometrics valuable.

The second and perhaps the most common use of econometrics is **hypothesis testing,** the evaluation of alternative theories with quantitative evidence. Much of economics involves building theoretical models and testing them against evidence, and hypothesis testing is vital to that scientific approach. For example, you could test the hypothesis that the product in Equation 1.1 is what economists call a normal good (one for which the quantity demanded increases when disposable income increases). You could do this by applying various statistical tests to the estimated coefficient (0.23) of disposable income (Yd) in Equation 1.2. At first glance, the evidence would seem to support this hypothesis because the coefficient's sign is positive, but the "statistical significance" of that estimate would have to be investigated before such a conclusion could be justified. Even though the estimated coefficient is positive, as expected, it may not be sufficiently different from zero to imply that the true coefficient is indeed positive instead of zero. Unfortunately, statistical tests of such hypotheses are not always easy, and there are times when two researchers can look at the same set of data and come to slightly different conclusions. Even given this possibility, the use of econometrics in testing hypotheses is probably its most important function.

The third and most difficult use of econometrics is to **forecast** or predict what is likely to happen next quarter, next year, or further into the future, based on what has happened in the past. For example, economists use econometric models to make forecasts of variables like sales, profits, Gross Domestic Product (GDP), and the inflation rate. The accuracy of such forecasts depends in large measure on the degree to which the past is a good guide to the future. Business leaders and politicians tend to be especially in-

3. The results in Equation 1.2 are from a model of the demand for chicken that we will examine in more detail in Section 6.1.

terested in this use of econometrics because they need to make decisions about the future, and the penalty for being wrong (bankruptcy for the entrepreneur and political defeat for the candidate) is high. To the extent that econometrics can shed light on the impact of their policies, business and government leaders will be better equipped to make decisions. For example, if the president of a company that sold the product modeled in Equation 1.1 wanted to decide whether to increase prices, forecasts of sales with and without the price increase could be calculated and compared to help make such a decision. In this way, econometrics can be used not only for forecasting but also for policy analysis.

1.1.2 Alternative Econometric Approaches

There are many different approaches to quantitative work. For example, the fields of biology, psychology, and physics all face quantitative questions similar to those faced in economics and business. However, these fields tend to use somewhat different techniques for analysis because the problems they face aren't the same. "We need a special field called econometrics, and textbooks about it, because it is generally accepted that economic data possess certain properties that are not considered in standard statistics texts or are not sufficiently emphasized there for use by economists."[4]

Different approaches also make sense within the field of economics. The kind of econometric tools used to quantify a particular function depends in part on the uses to which that equation will be put. A model built solely for descriptive purposes might be different from a forecasting model, for example.

To get a better picture of these approaches, let's look at the steps necessary for any kind of quantitative research:

1. specifying the models or relationships to be studied
2. collecting the data needed to quantify the models
3. quantifying the models with the data

Steps 1 and 2 are similar in all quantitative work, but the techniques used in step 3, quantifying the models, differ widely between and within disciplines. Choosing the best technique for a given model is a theory-based skill that is often referred to as the "art" of econometrics. There are many alternative approaches to quantifying the same equation, and each approach may

4. Clive Granger, "A Review of Some Recent Textbooks of Econometrics," *Journal of Economic Literature*, March 1994, p. 117.

give somewhat different results. The choice of approach is left to the individual econometrician (the researcher using econometrics), but each researcher should be able to justify that choice.

This book will focus primarily on one particular econometric approach: *single-equation linear regression analysis.* The majority of this book will thus concentrate on regression analysis, but it is important for every econometrician to remember that regression is only one of many approaches to econometric quantification.

The importance of critical evaluation cannot be stressed enough; a good econometrician can diagnose faults in a particular approach and figure out how to repair them. The limitations of the regression analysis approach must be fully perceived and appreciated by anyone attempting to use regression analysis or its findings. The possibility of missing or inaccurate data, incorrectly formulated relationships, poorly chosen estimating techniques, or improper statistical testing procedures implies that the results from regression analyses should always be viewed with some caution.

1.2 What Is Regression Analysis?

Econometricians use regression analysis to make quantitative estimates of economic relationships that previously have been completely theoretical in nature. After all, anybody can claim that the quantity of compact discs demanded will increase if the price of those discs decreases (holding everything else constant), but not many people can put specific numbers into an equation and estimate *by how many* compact discs the quantity demanded will increase for each dollar that price decreases. To predict the *direction* of the change, you need a knowledge of economic theory and the general characteristics of the product in question. To predict the *amount* of the change, though, you need a sample of data, and you need a way to estimate the relationship. The most frequently used method to estimate such a relationship in econometrics is regression analysis.

1.2.1 Dependent Variables, Independent Variables, and Causality

Regression analysis is a statistical technique that attempts to "explain" movements in one variable, the **dependent variable,** as a function of movements in a set of other variables, called the **independent** (or **explanatory**) **variables,** through the quantification of a single equation. For example in Equation 1.1:

$$Q = f(P, P_s, Yd) \tag{1.1}$$

Q is the dependent variable and P, P_s, and Yd are the independent variables. Regression analysis is a natural tool for economists because most (though not all) economic propositions can be stated in such single-equation functional forms. For example, the quantity demanded (dependent variable) is a function of price, the prices of substitutes, and income (independent variables).

Much of economics and business is concerned with cause-and-effect propositions. If the price of a good increases by one unit, then the quantity demanded decreases on average by a certain amount, depending on the price elasticity of demand (defined as the percentage change in the quantity demanded that is caused by a one percent change in price). Similarly, if the quantity of capital employed increases by one unit, then output increases by a certain amount, called the marginal productivity of capital. Propositions such as these pose an if-then, or causal, relationship that logically postulates that a dependent variable's movements are causally determined by movements in a number of specific independent variables.

Don't be deceived by the words dependent and independent, however. Although many economic relationships are causal by their very nature, a regression result, no matter how statistically significant, cannot prove causality. All regression analysis can do is test whether a significant quantitative relationship exists. Judgments as to causality must also include a healthy dose of economic theory and common sense. For example, the fact that the bell on the door of a flower shop rings just before a customer enters and purchases some flowers by no means implies that the bell causes purchases! If events A and B are related statistically, it may be that A causes B, that B causes A, that some omitted factor causes both, or that a chance correlation exists between the two.

The cause-and-effect relationship is often so subtle that it fools even the most prominent economists. For example, in the late nineteenth century, English economist Stanley Jevons hypothesized that sunspots caused an increase in economic activity. To test this theory, he collected data on national output (the dependent variable) and sunspot activity (the independent variable) and showed that a significant positive relationship existed. This result led him, and some others, to jump to the conclusion that sunspots did indeed cause output to rise. Such a conclusion was unjustified because regression analysis cannot confirm causality; it can only test the strength and direction of the quantitative relationships involved.

1.2.2 Single-Equation Linear Models

The simplest single-equation linear regression model is:

$$Y = \beta_0 + \beta_1 X \qquad (1.3)$$

Equation 1.3 states that Y, the dependent variable, is a single-equation linear function of X, the independent variable. The model is a single-equation model because no equation for X as a function of Y (or any other variable) has been specified. The model is linear because if you were to plot Equation 1.3 on graph paper, it would be a straight line rather than a curve.

The βs are the **coefficients** that determine the coordinates of the straight line at any point. β_0 is the **constant** or **intercept** term; it indicates the value of Y when X equals zero. β_1 is the **slope coefficient,** and it indicates the amount that Y will change when X increases by one unit. The solid line in Figure 1.1 illustrates the relationship between the coefficients and the graphical meaning of the regression equation. As can be seen from the diagram, Equation 1.3 is indeed linear.

The slope coefficient, β_1, shows the response of Y to a change in X. Since being able to explain and predict changes in the dependent variable is the essential reason for quantifying behavioral relationships, much of the emphasis in regression analysis is on slope coefficients such as β_1. In Figure 1.1 for example, if X were to increase from X_1 to X_2 (ΔX), the value of Y in Equation 1.3 would increase from Y_1 to Y_2 (ΔY). For linear (i.e., straight-line) regression models, the response in the predicted value of Y due to a change in X is constant and equal to the slope coefficient β_1:

$$\frac{(Y_2 - Y_1)}{(X_2 - X_1)} = \frac{\Delta Y}{\Delta X} = \beta_1$$

where Δ is used to denote a change in the variables. Some readers may recognize this as the "rise" (ΔY) divided by the "run" (ΔX). For a linear model, the slope is constant over the entire function.

We must distinguish between an equation that is linear in the variables and one that is linear in the coefficients. This distinction is important because if linear regression techniques are going to be applied to an equation, that equation *must be* linear in the coefficients.

An equation is **linear in the variables** if plotting the function in terms of X and Y generates a straight line. For example, Equation 1.3:

$$Y = \beta_0 + \beta_1 X \tag{1.3}$$

is linear in the variables, but Equation 1.4:

$$Y = \beta_0 + \beta_1 X^2 \tag{1.4}$$

is not linear in the variables because if you were to plot Equation 1.4 it

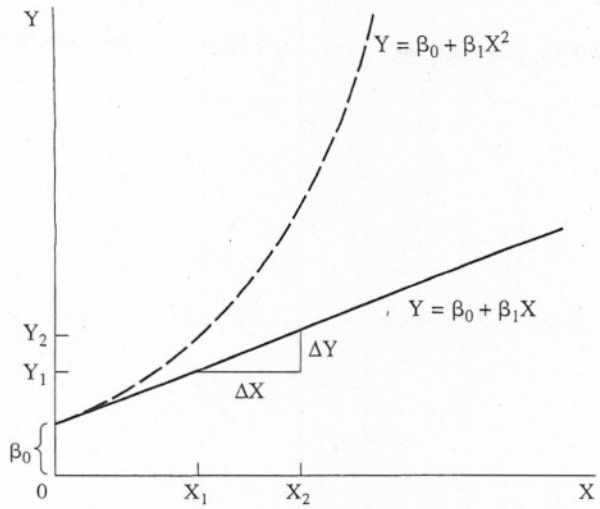

Figure 1.1 Graphical Representation of the Coefficients of the Regression Line

The graph of the equation $Y = \beta_0 + \beta_1 X$ is linear with a constant slope equal to $\beta_1 = \Delta Y/\Delta X$. The graph of the equation $Y = \beta_0 + \beta_1 X^2$, on the other hand, is nonlinear with an increasing slope (if $\beta_1 > 0$).

would be a quadratic, not a straight line. This difference[5] can be seen in Figure 1.1.

An equation is **linear in the coefficients** only if the coefficients (the βs) appear in their simplest form—they are not raised to any powers (other than one), are not multiplied or divided by other coefficients, and do not themselves include some sort of function (like logs or exponents). For example, Equation 1.3 is linear in the coefficients, but Equation 1.5:

$$Y = \beta_0 + X^{\beta_1} \tag{1.5}$$

is not linear in the coefficients β_0 and β_1. Equation 1.5 is not linear because there is no rearrangement of the equation that will make it linear in the βs of original interest, β_0 and β_1. In fact, of all possible equations for a single explanatory variable, *only* functions of the general form:

$$f(Y) = \beta_0 + \beta_1 f(X) \tag{1.6}$$

5. Equations 1.3 and 1.4 have the same β_0 in Figure 1.1 for comparison purposes only. If the equations were applied to the same data, the estimated β_0s would be different.

are linear in the coefficients β_0 and β_1. In essence, any sort of configuration of the Xs and Ys can be used and the equation will continue to be linear in the coefficients. However, even a slight change in the configuration of the βs will cause the equation to become nonlinear in the coefficients.

Although linear regressions need to be linear in the coefficients, they do not necessarily need to be linear in the variables. Linear regression analysis can be applied to an equation that is nonlinear in the variables if the equation can be formulated in a way that is linear in the coefficients. Indeed, when econometricians use the phrase "linear regression," they usually mean "regression that is linear in the coefficients."[6]

1.2.3 The Stochastic Error Term

Besides the variation in the dependent variable (Y) that is caused by the independent variable (X), there is almost always variation that comes from other sources as well. This additional variation comes in part from omitted explanatory variables (e.g., X_2 and X_3). However, even if these extra variables are added to the equation, there still is going to be some variation in Y that simply cannot be explained by the model.[7] This variation probably comes from sources such as omitted influences, measurement error, incorrect functional form, or purely random and totally unpredictable occurrences. By *random* we mean something that has its value determined entirely by chance.

Econometricians admit the existence of such inherent unexplained variation ("error") by explicitly including a stochastic (or random) error term in their regression models. A **stochastic error term** is a term that is added to a regression equation to introduce all of the variation in Y that cannot be explained by the included Xs. It is, in effect, a symbol of the econometrician's ignorance or inability to model all the movements of the dependent variable.

6. The application of regression analysis to equations that are nonlinear in the variables is covered in Chapter 7. The application of regression techniques to equations that are nonlinear in the *coefficients*, however, is much more difficult.

7. The exception would be the extremely rare case where the data can be explained by some sort of physical law and are measured perfectly. Here, continued variation would point to an omitted independent variable. A similar kind of problem is often encountered in astronomy, where planets can be discovered by noting that the orbits of known planets exhibit variations that can be caused only by the gravitational pull of another heavenly body. Absent these kinds of physical laws, researchers in economics and business would be foolhardy to believe that *all* variation in Y can be explained by a regression model because there are always elements of error in any attempt to measure a behavioral relationship.

The error term (sometimes called a disturbance term) is usually referred to with the symbol epsilon (ϵ), although other symbols (like u or v) are sometimes used.

The addition of a stochastic error term (ϵ) to Equation 1.3 results in a typical regression equation:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1.7}$$

Equation 1.7 can be thought of as having two components, the *deterministic* component and the *stochastic,* or random, component. The expression $\beta_0 + \beta_1 X$ is called the *deterministic* component of the regression equation because it indicates the value of Y that is determined by a given value of X, which is assumed to be nonstochastic. This deterministic component can also be thought of as the **expected value** of Y given X, the mean value of the Ys associated with a particular value of X. For example, if the average height of all 14-year-old girls is 5 feet, then 5 feet is the expected value of a girl's height given that she is 14. The deterministic part of the equation may be written:

$$E(Y|X) = \beta_0 + \beta_1 X \tag{1.8}$$

which states that the expected value of Y given X, denoted as $E(Y|X)$, is a linear function of the independent variable (or variables if there are more than one).[8]

Unfortunately, the value of Y observed in the real world is unlikely to be exactly equal to the deterministic expected value $E(Y|X)$. After all, not all 14-year-old girls are 5 feet tall. As a result, the stochastic element (ϵ) must be added to the equation:

$$Y = E(Y|X) + \epsilon = \beta_0 + \beta_1 X + \epsilon \tag{1.9}$$

8. This property holds as long as $E(\epsilon|X) = 0$ [read as "the expected value of X, given epsilon" equals zero], which is true as long as the Classical Assumptions (to be outlined in Chapter 4) are met. It's easiest to think of $E(\epsilon)$ as the mean of ϵ, but the expected value operator E technically is a summation of all the values that a function can take, weighted by the probability of each value. The expected value of a constant is that constant, and the expected value of a sum of variables equals the sum of the expected values of those variables.

The stochastic error term must be present in a regression equation because there are at least four sources of variation in Y other than the variation in the included Xs:

1. Many minor influences on Y are *omitted* from the equation (for example, because data are unavailable).
2. It is virtually impossible to avoid some sort of *measurement error* in at least one of the equation's variables.
3. The underlying theoretical equation might have a *different functional form* (or shape) than the one chosen for the regression. For example, the underlying equation might be nonlinear in the variables for a linear regression.
4. All attempts to generalize human behavior must contain at least some amount of unpredictable or *purely random* variation.

To get a better feeling for these components of the stochastic error term, let's think about a consumption function (aggregate consumption as a function of aggregate disposable income). First, consumption in a particular year may have been less than it would have been because of uncertainty over the future course of the economy. Since this uncertainty is hard to measure, there might be no variable measuring consumer uncertainty in the equation. In such a case, the impact of the omitted variable (consumer uncertainty) would likely end up in the stochastic error term. Second, the observed amount of consumption may have been different from the actual level of consumption in a particular year due to an error (such as a sampling error) in the measurement of consumption in the National Income Accounts. Third, the underlying consumption function may be nonlinear, but a linear consumption function might be estimated. (To see how this incorrect functional form would cause errors, see Figure 1.2.) Fourth, the consumption function attempts to portray the behavior of people, and there is always an element of unpredictability in human behavior. At any given time, some random event might increase or decrease aggregate consumption in a way that might never be repeated and couldn't be anticipated.

These possibilities explain the existence of a difference between the observed values of Y and the values expected from the deterministic component of the equation, $E(Y|X)$. These sources of error will be covered in more detail in the following chapters, but for now it is enough to recognize that in econometric research there will always be some stochastic or random element, and, for this reason, an error term must be added to all regression equations.

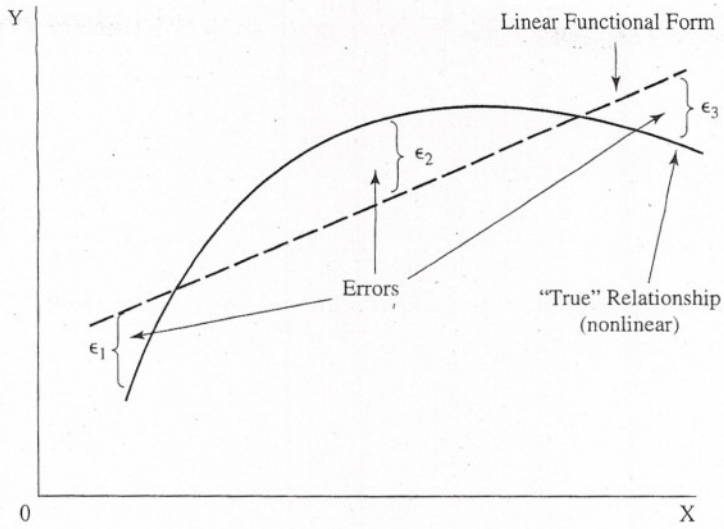

Figure 1.2 Errors Caused by Using a Linear Functional Form to Model a Nonlinear Relationship

One source of stochastic error is the use of an incorrect functional form. For example, if a linear functional form is used when the underlying relationship is nonlinear, systematic errors (the ϵs) will occur. These nonlinearities are just one component of the stochastic error term. The others are omitted variables, measurement error, and purely random variation.

1.2.4 Extending the Notation

Our regression notation needs to be extended to include reference to the number of observations and to allow the possibility of more than one independent variable. If we include a specific reference to the observations, the single-equation linear regression model may be written as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (i = 1, 2, \ldots, n) \qquad (1.10)$$

where:
- Y_i = the ith observation[9] of the dependent variable
- X_i = the ith observation of the independent variable
- ϵ_i = the ith observation of the stochastic error term
- β_0, β_1 = the regression coefficients
- n = the number of observations

9. A typical observation (or unit of analysis) is an individual person, year, or country. For example, a series of annual observations starting in 1950 would have Y_1 = Y for 1950, Y_2 for 1951, etc.

This equation is actually n equations, one for each of the n observations:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \epsilon_2 \\ Y_3 &= \beta_0 + \beta_1 X_3 + \epsilon_3 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_n + \epsilon_n \end{aligned}$$

That is, the regression model is assumed to hold for each observation. The coefficients do not change from observation to observation, but the values of Y, X, and ϵ do.

A second notational addition allows for more than one independent variable. Since more than one independent variable is likely to have an effect on the dependent variable, our notation should allow these additional explanatory Xs to be added. If we define:

X_{1i} = the *i*th observation of the first independent variable
X_{2i} = the *i*th observation of the second independent variable
X_{3i} = the *i*th observation of the third independent variable

then all three variables can be expressed as determinants of Y in a **multivariate** (more than one independent variable) linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \qquad (1.11)$$

The **meaning of the regression coefficient** β_1 in this equation is the impact of a one unit increase in X_1 on the dependent variable Y, *holding constant* the other included independent variables (X_2 and X_3). Similarly, β_2 gives the impact of a one-unit increase in X_2 on Y, holding X_1 and X_3 constant. These **multivariate regression coefficients** (which are parallel in nature to partial derivatives in calculus) serve to isolate the impact on Y of a change in one variable from the impact on Y of changes in the other variables. This is possible because multivariate regression takes the movements of X_2 and X_3 into account when it estimates the coefficient of X_1. The result is quite similar to what we would obtain if we were capable of conducting controlled laboratory experiments in which only one variable at a time was changed.

In the real world, though, it is almost impossible to run controlled experiments, because many economic factors change simultaneously, often in opposite directions. Thus the ability of regression analysis to measure the impact of one variable on the dependent variable, *holding constant the influence of the other variables in the equation,* is a tremendous advantage. Note that if a variable is not included in an equation, then its impact is *not* held constant in the estimation of the regression coefficients. This will be discussed further in Chapter 6.

The general multivariate regression model with K independent variables thus is written as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i \quad (i = 1, 2, \ldots, n) \tag{1.12}$$

If the sample consists of a series of years or months (called a **time series**), then the subscript i is usually replaced with a t to denote time.[10]

1.3 The Estimated Regression Equation

Once a specific equation has been decided upon, it must be quantified. This quantified version of the theoretical regression equation is called the **estimated regression equation** and is obtained from a sample of actual Xs and Ys. Although the theoretical equation is purely abstract in nature:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1.13}$$

the estimated regression equation has actual numbers in it:

$$\hat{Y}_i = 103.40 + 6.38 X_i \tag{1.14}$$

The observed, real-world values of X and Y are used to calculate the coefficient estimates 103.40 and 6.38. These estimates are used to determine $\hat{Y}$ (read as "Y-hat"), the *estimated* or *fitted* value of Y.

Let's look at the differences between a theoretical regression equation and an estimated regression equation. First, the theoretical regression coefficients β_0 and β_1 in Equation 1.13 have been replaced with *estimates* of those coefficients like 103.40 and 6.38 in Equation 1.14. We can't actually observe the values of the true[11] regression coefficients, so instead we calculate estimates of those coefficients from the data. The **estimated regression coefficients,**

10. It also does not matter if X_{1i}, for example, is written as X_{i1} as long as the appropriate definitions are presented. Often the observational subscript (i or t) is deleted, and the reader is expected to understand that the equation holds for each observation in the sample.

11. Our use of the word *true* throughout the text should be taken with a grain of salt. Many philosophers argue that the concept of truth is useful only relative to the scientific research program in question. Many economists agree, pointing out that what is true for one generation may well be false for another. To us, the true coefficient is the one that you'd obtain if you could run a regression on the entire relevant population. Thus, readers who so desire can substitute the phrase "population coefficient" for "true coefficient" with no loss in meaning.

more generally denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ (read as "beta-hats"), are empirical best guesses of the true regression coefficients and are obtained from data from a sample of the Ys and Xs. The expression

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{1.15}$$

is the empirical counterpart of the theoretical regression Equation 1.13. The calculated estimates in Equation 1.14 are examples of estimated regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$. For each sample we calculate a different set of estimated regression coefficients.

$\hat{Y}_i$ is the *estimated value* of Y_i, and it represents the value of Y calculated from the estimated regression equation for the *i*th observation. As such, $\hat{Y}_i$ is our predication of $E(Y_i|X_i)$ from the regression equation. The closer $\hat{Y}_i$ is to Y_i, the better the fit of the equation. (The word *fit* is used here much as it would be used to describe how well clothes fit.)

The difference between the estimated value of the dependent variable ($\hat{Y}_i$) and the actual value of the dependent variable (Y_i) is defined as the **residual** (e_i):

$$e_i = Y_i - \hat{Y}_i \tag{1.16}$$

Note the distinction between the residual in Equation 1.16 and the error term:

$$\epsilon_i = Y_i - E(Y_i|X_i) \tag{1.17}$$

The *residual* is the difference between the observed Y and the estimated regression line ($\hat{Y}$), while the *error term* is the difference between the observed Y and the true regression equation (the expected value of Y). Note that the error term is a theoretical concept that can never be observed, but the residual is a real-world value that is calculated for each observation every time a regression is run. Most regression techniques not only calculate the residuals but also attempt to select values of $\hat{\beta}_0$ and $\hat{\beta}$ that keep the residuals as low as possible. The smaller the residuals, the better the fit, and the closer the $\hat{Y}$s will be to the Ys.

All these concepts are shown in Figure 1.3. The (X, Y) pairs are shown as points on the diagram, and both the true regression equation (which cannot be seen in real applications) and an estimated regression equation are included. Notice that the estimated equation is close to but not equivalent to the true line. This is a typical result. For example, $\hat{Y}_6$, the computed value of Y

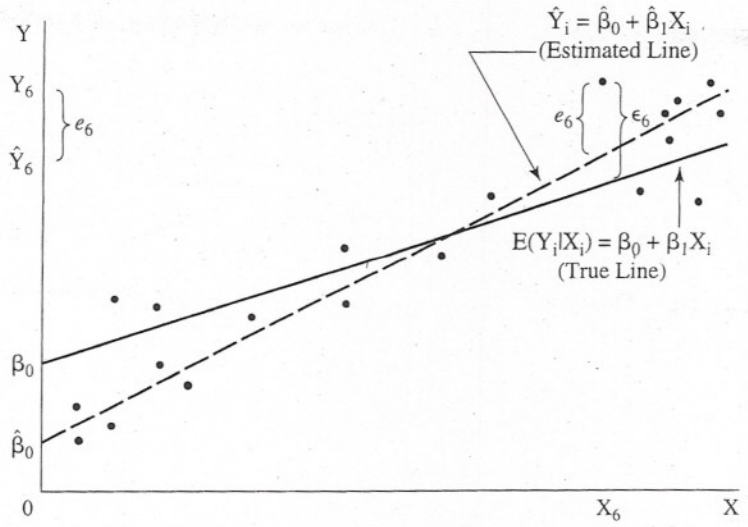

Figure 1.3 True and Estimated Regression Lines

The true relationship between X and Y (the solid line) cannot typically be observed, but the estimated regression line (the dotted line) can. The difference between an observed data point (for example, i = 6) and the true line is the value of the stochastic error term (ϵ_6). The difference between the observed Y_6 and the estimated value from the regression line ($\hat{Y}_6$) is the value of the residual for this observation, e_6.

for the sixth observation, lies on the estimated (dashed) line, and it differs from Y_6, the actual observed value of Y for the sixth observation. The difference between the observed and estimated values is the residual, denoted by e_6. In addition, although we usually would not be able to see an observation of the error term, we have drawn the assumed true regression line here (the solid line) to see the sixth observation of the error term, ϵ_6, which is the difference between the true line and the observed value of Y, Y_6.

Another way to state the estimated regression equation is to combine Equations 1.15 and 1.16, obtaining:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \tag{1.18}$$

Compare this equation to Equation 1.13. When we replace the theoretical regression coefficients with estimated coefficients, the error term must be replaced by the residual, because the error term, like the regression coefficients β_0 and β_1, can never be observed. Instead, the residual is observed and measured whenever a regression line is estimated with a sample of Xs and Ys. In

this sense, the residual can be thought of as an estimate of the error term, and e could have been denoted as $\hat{\epsilon}$.

The following chart summarizes the notation used in the true and estimated regression equations:

True Regression Equation	Estimated Regression Equation
β_0	$\hat{\beta}_0$
β_1	$\hat{\beta}_1$
ϵ_i	e_i

The estimated regression model can be extended to more than one independent variable by adding the additional Xs to the right side of the equation. The multivariate estimated regression counterpart of Equation 1.12 is:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki} \qquad (1.19)$$

1.4 A Simple Example of Regression Analysis

Let's look at a fairly simple example of regression analysis. Suppose you've accepted a summer job as a weight guesser at the local amusement park, Magic Hill. Customers pay 50 cents each, which you get to keep if you guess their weight within 10 pounds. If you miss by more than 10 pounds, then you have to give the customer a small prize that you buy from Magic Hill for 60 cents each. Luckily, the friendly managers of Magic Hill have arranged a number of marks on the wall behind the customer so that you are capable of measuring the customer's height accurately. Unfortunately, there is a five-foot wall between you and the customer, so you can tell little about the person except for height and (usually) gender.

On your first day on the job, you do so poorly that you work all day and somehow manage to lose two dollars, so on the second day you decide to collect data to run a regression to estimate the relationship between weight and height. Since most of the participants are male, you decide to limit your sample to males. You hypothesize the following theoretical relationship:

$$Y_i = f(\overset{+}{X}_i) + \epsilon_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (1.20)$$

where: Y_i = the weight (in pounds) of the *i*th customer
X_i = the height (in inches above 5 feet) of the *i*th customer
ϵ_i = the value of the stochastic error term for the *i*th customer

TABLE 1.1 DATA FOR AND RESULTS OF THE WEIGHT-GUESSING EQUATION

Observation i (1)	Height Above 5′ X_i (2)	Weight Y_i (3)	Predicted Weight $\hat{Y}_i$ (4)	Residual e_i (5)	$ Gain or Loss (6)
1	5.0	140.0	135.3	4.7	+.50
2	9.0	157.0	160.8	−3.8	+.50
3	13.0	205.0	186.3	18.7	−.60
4	12.0	198.0	179.9	18.1	−.60
5	10.0	162.0	167.2	−5.2	+.50
6	11.0	174.0	173.6	0.4	+.50
7	8.0	150.0	154.4	−4.4	+.50
8	9.0	165.0	160.8	4.2	+.50
9	10.0	170.0	167.2	2.8	+.50
10	12.0	180.0	179.9	0.1	+.50
11	11.0	170.0	173.6	−3.6	+.50
12	9.0	162.0	160.8	1.2	+.50
13	10.0	165.0	167.2	−2.2	+.50
14	12.0	180.0	179.9	0.1	+.50
15	8.0	160.0	154.4	5.6	+.50
16	9.0	155.0	160.8	−5.8	+.50
17	10.0	165.0	167.2	−2.2	+.50
18	15.0	190.0	199.1	−9.1	+.50
19	13.0	185.0	186.3	−1.3	+.50
20	11.0	155.0	173.6	−18.6	−.60
				TOTAL =	$6.70

Note: This data set, and every other data set in the text, is available on the text's website in four formats and on the EViews CD-ROM. This data set's filename is HTWT1

In this case, the sign of the theoretical relationship between height and weight is believed to be positive (signified by the positive sign above X_i in the general theoretical equation), but you must quantify that relationship in order to estimate weights given heights. To do this, you need to collect a data set, and you need to apply regression analysis to the data.

The next day you collect the data summarized in Table 1.1 and run your regression on the Magic Hill computer, obtaining the following estimates:

$$\hat{\beta}_0 = 103.40 \qquad \hat{\beta}_1 = 6.38$$

This means that the equation

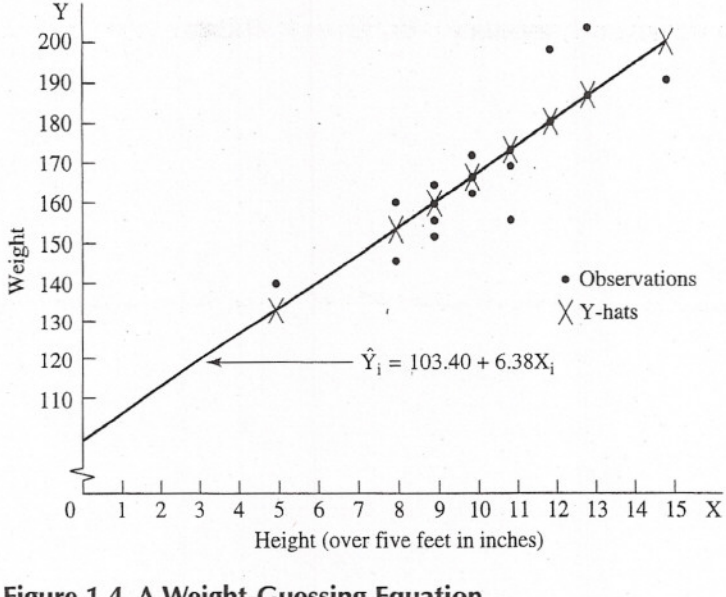

Figure 1.4 A Weight-Guessing Equation

If we plot the data from the weight-guessing example and include the estimated regression line, we can see that the estimated $\hat{Y}$s come fairly close to the observed Ys for all but three observations. Find a male friend's height and weight on the graph; how well does the regression equation work?

$$\text{Estimated weight} = 103.40 + 6.38 \cdot \text{Height (inches above five feet)} \tag{1.21}$$

is worth trying as an alternative to just guessing the weights of your customers. Such an equation estimates weight with a constant base of 103.40 pounds and adds 6.38 pounds for every inch of height over 5 feet. Note that the sign of $\hat{\beta}_1$ is positive, as you expected.

How well does the equation work? To answer this question, you need to calculate the residuals (Y_i minus $\hat{Y}_i$) from Equation 1.21 to see how many were greater than ten. As can be seen in the last column in Table 1.1, if you had applied the equation to these 20 people you wouldn't exactly have gotten rich, but at least you would have earned $6.70 instead of losing $2.00. Figure 1.4 shows not only Equation 1.21 but also the weight and height data for all 20 customers used as the sample.

Equation 1.21 would probably help a beginning weight guesser, but it could be improved by adding other variables or by collecting a larger sample. Such an equation is realistic, though, because it's likely that every successful

weight guesser uses an equation like this without consciously thinking about that concept.

Our goal with this equation was to quantify the theoretical weight/height equation, Equation 1.20, by collecting data (Table 1.1) and calculating an estimated regression, Equation 1.21. Although the true equation, like observations of the stochastic error term, can never be known, we were able to come up with an estimated equation that had the sign we expected for $\hat{\beta}_1$ and that helped us in our job. Before you decide to quit school or your job and try to make your living guessing weights at Magic Hill, there is quite a bit more to learn about regression analysis, so we'd better move on.

1.5 Using Regression to Explain Housing Prices

As much fun as guessing weights at an amusement park might be, it's hardly a typical example of the use of regression analysis. For every regression run on such an off-the-wall topic, there are literally hundreds run to *describe* the reaction of GDP to an increase in the money supply, to *test* an economic theory with new data, or to *forecast* the effect of a price change on a firm's sales.

As a more realistic example, let's look at a model of housing prices. The purchase of a house is probably the most important financial decision in an individual's life, and one of the key elements in that decision is an appraisal of the house's value. If you overvalue the house, you can lose thousands of dollars by paying too much; if you undervalue the house, someone might outbid you.

All this wouldn't be much of a problem if houses were homogeneous products, like corn or gold, that have generally known market prices with which to compare a particular asking price. Such is hardly the case in the real estate market. Consequently, an important element of every housing purchase is an appraisal of the market value of the house, and many real estate appraisers use regression analysis to help them in their work.

Suppose your family is about to buy a house in Southern California, but you're convinced that the owner is asking too much money. The owner says that the asking price of $230,000 is fair because a larger house next door sold for $230,000 about a year ago. You're not sure it's reasonable to compare the prices of different-sized houses that were purchased at different times. What can you do to help decide whether to pay the $230,000?

Since you're taking an econometrics class, you decide to collect data on all local houses that were sold within the last few weeks and to build a re-

gression model of the sales prices of the houses as a function of their sizes.[12] Such a data set is called **cross-sectional** because all of the observations are from the same point in time and represent different individual economic entities (like countries, or in this case, houses) from that same point in time.

To measure the impact of size on price, you include the size of the house as an independent variable in a regression equation that has the price of that house as the dependent variable. You expect a positive sign for the coefficient of size, since big houses cost more to build and tend to be more desirable than small ones. Thus the theoretical model is:

$$P_i = f(\overset{+}{S_i}) + \epsilon_i = \beta_0 + \beta_1 S_i + \epsilon_i \qquad (1.22)$$

where: P_i = the price (in thousands of \$) of the ith house
S_i = the size (in square feet) of that house
ϵ_i = the value of the stochastic error term for that house

You collect the records of all recent real estate transactions, find that 43 local houses were sold within the last 4 weeks, and estimate the following regression of those 43 observations:

$$\hat{P}_i = 40.0 + 0.138 S_i \qquad (1.23)$$

What do these estimated coefficients mean? The most important coefficient is $\hat{\beta}_1 = 0.138$, since the reason for the regression is to find out the impact of size on price. This coefficient means that if size increases by 1 square foot, price will increase by 0.138 thousand dollars (\$138). $\hat{\beta}_1$ thus measures the change in P_i associated with a one-unit increase in S_i. It's the slope of the regression line in a graph like Figure 1.5.

What does $\hat{\beta}_0 = 40.0$ mean? $\hat{\beta}_0$ is the estimate of the constant or intercept term. In our equation, it means that price equals 40.0 when size equals zero. As can be seen in Figure 1.5, the estimated regression line intersects the price axis at 40.0. While it might be tempting to say that the average price of a vacant lot is \$40,000, such a conclusion would be unjustified for a number of

12. It's unusual for an economist to build a model of price without including some measure of quantity on the right-hand side. Such models of the price of a good as a function of the attributes of that good are called *hedonic* models and will be discussed in greater depth in Section 11.7. The interested reader is encouraged to skim the first few paragraphs of that section before continuing on with this example.

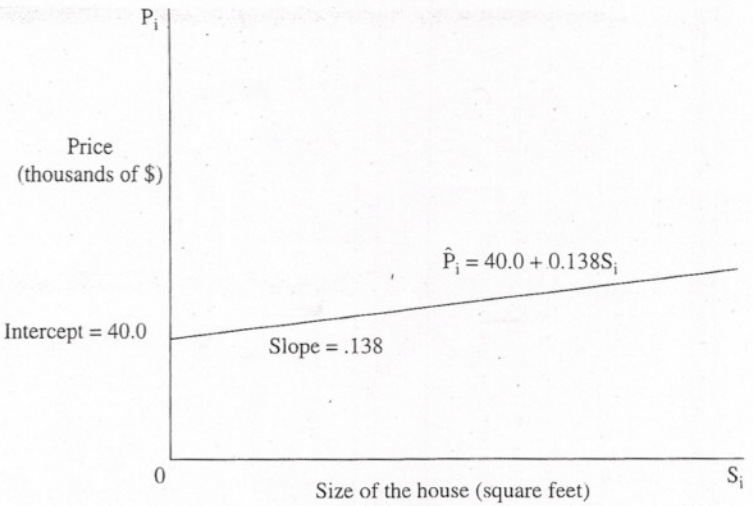

Figure 1.5 A Cross Sectional Model of Housing Prices

A regression equation that has the price of a house in Southern California as a function of the size of that house has an intercept of 40.0 and a slope of 0.138, using Equation 1.23.

reasons, which will be discussed in later chapters. It's much safer either to interpret $\hat{\beta}_0 = 40.0$ as nothing more than the value of the estimated regression when $S_i = 0$, or to not interpret $\hat{\beta}_0$ at all.

How can you use this estimated regression to help decide whether to pay $230,000 for the house? If you calculate a $\hat{Y}$ (predicted price) for a house that is the same size (1,600 square feet) as the one you're thinking of buying, you can then compare this $\hat{Y}$ with the asking price of $230,000. To do this, substitute 1600 for S_i in Equation 1.23, obtaining:

$$\hat{P}_i = 40.0 + 0.138(1600) = 40.0 + 220.8 = 260.8$$

The house seems to be a good deal. The owner is asking "only" $230,000 for a house when the size implies a price of $260,800! Perhaps your original feeling that the price was too high was a reaction to the steep housing prices in Southern California in general and not a reflection of this specific price.

On the other hand, perhaps the price of a house is influenced by more than just the size of the house. (After all, what good's a house in Southern California unless it has a pool or air-conditioning?) Such multivariate models are the heart of econometrics, but we'll hold off adding more indepen-

dent variables to Equation 1.23 until we return to this housing price example later in the text.

1.6 Summary

1. Econometrics, literally "economic measurement," is a branch of economics that attempts to quantify theoretical relationships. Regression analysis is only one of the techniques used in econometrics, but it is by far the most frequently used.
2. The major uses of econometrics are description, hypothesis testing, and forecasting. The specific econometric techniques employed may vary depending on the use of the research.
3. While regression analysis specifies that a dependent variable is a function of one or more independent variables, regression analysis alone cannot prove or even imply causality.
4. Linear regression can only be applied to equations that are *linear in the coefficients,* which means that the regression coefficients are in their simplest possible form. For an equation with two explanatory variables, this form would be:

$$f(Y_i) = \beta_0 + \beta_1 f(X_{1i}) + \beta_2 f(X_{2i}) + \epsilon_i$$

5. A stochastic error term must be added to all regression equations to account for variations in the dependent variable that are not explained completely by the independent variables. The components of this error term include:
   a. omitted or left-out variables
   b. measurement errors in the data
   c. an underlying theoretical equation that has a different functional form (shape) than the regression equation
   d. purely random and unpredictable events
6. An estimated regression equation is an approximation of the true equation that is obtained by using data from a sample of actual Ys and Xs. Since we can never know the true equation, econometric analysis focuses on this estimated regression equation and the estimates of the regression coefficients. The difference between a particular observation of the dependent variable and the value estimated from the regression equation is called the residual.

Exercises

(Answers to even-numbered exercises are in Appendix A.)

1. Write the meaning of each of the following terms without referring to the book (or your notes), and compare your definition with the version in the text for each:
   a. stochastic error term
   b. regression analysis
   c. linear in the variables
   d. slope coefficient
   e. multivariate regression model
   f. expected value
   g. residual
   h. linear in the coefficients

2. Use your own computer's regression software and the weight (Y) and height (X) data from Table 1.1 to see if you can reproduce the estimates in Equation 1.21. There are three different ways to load the data: You can type in the data yourself, you can open datafile HTWT1 on the EViews CD, or you can download datafile HTWT1 (in any of four formats: SAS, EXCEL, SHAZAM, and ASCII) from the text's website: www.awlonline.com/studenmund/ Once the datafile is loaded, then run Y = f(X), and your results should match Equation 1.21. Different programs require different commands to run a regression. For help in how to do this with EViews, for example, see the answer to this question in Appendix A.

3. Decide whether you would expect relationships between the following pairs of dependent and independent variables (respectively) to be positive, negative, or ambiguous. Explain your reasoning.
   a. Aggregate net investment in the U.S. in a given year and GDP in that year.
   b. The amount of hair on the head of a male professor and the age of that professor.
   c. The number of acres of wheat planted in a season and the price of wheat at the beginning of that season.
   d. Aggregate net investment and the real rate of interest in the same year and country.
   e. The growth rate of GDP in a year and the average hair length in that year.
   f. The quantity of canned heat demanded and the price of a can of heat.

4. Let's return to the height/weight example in Section 1.4:
   a. Go back to the data set and identify the three customers who seem to be quite a distance from the estimated regression line. Would we have a better regression equation if we dropped these customers from the sample?
   b. Measure the height of a male friend and plug it into Equation 1.21. Does the equation come within ten pounds? If not, do you think you see why? Why does the estimated equation predict the same weight for all males of the same height when it is obvious that all males of the same height don't weigh the same?
   c. Look over the sample with the thought that it might not be randomly drawn. Does the sample look abnormal in any way? (*Hint:* Are the customers who choose to play such a game a random sample?) If the sample isn't random, would this have an effect on the regression results and the estimated weights?
   d. Think of at least one other factor besides height that might be a good choice as a variable in the weight/height equation. How would you go about obtaining the data for this variable? What would the expected sign of your variable's coefficient be if the variable were added to the equation?

5. Continuing with the height/weight example, suppose you collected data on the heights and weights of 29 more customers and estimated the following equation:

$$\hat{Y}_i = 125.1 + 4.03X_i \qquad (1.24)$$

   where: Y_i = the weight (in pounds) of the ith person
   X_i = the height (in inches over five feet) of the ith person

   a. Why aren't the coefficients in Equation 1.24 the same as those we estimated previously (Equation 1.21)?
   b. Compare the estimated coefficients of Equation 1.24 with those in Equation 1.21. Which equation has the steeper estimated relationship between height and weight? Which equation has the higher intercept? At what point do the two intersect?
   c. Use Equation 1.24 to "predict" the 20 original weights given the heights in Table 1.1. How many weights does Equation 1.24 miss by more than ten pounds? Does Equation 1.24 do better or worse than Equation 1.21? Could you have predicted this result beforehand?
   d. Suppose you had one last day on the weight-guessing job. What equation would you use to guess weights? (*Hint:* There is more than one possible answer.)

6. Not all regression coefficients have positive expected signs. For example, a *Sports Illustrated* article by Jaime Diaz reported on a study of golfing putts of various lengths on the Professional Golfers Association (PGA) Tour.[13] The article included data on the percentage of putts made (P_i) as a function of the length of the putt in feet (L_i). Since the longer the putt, the less likely even a professional is to make it, we'd expect L_i to have a negative coefficient in an equation explaining P_i. Sure enough, if you estimate an equation on the data in the article, you obtain:

$$\hat{P}_i = f(\overline{L}_i) = 83.6 - 4.1L_i \qquad (1.25)$$

   a. Carefully write out the exact meaning of the coefficient of L_i.
   b. Use Equation 1.25 to determine the percent of the time you'd expect a PGA golfer to make a 10-foot putt. Does this seem realistic? How about a 1-foot putt or a 25-foot putt? Do these seem as realistic?
   c. Your answer to part b should suggest that there's a problem in applying a linear regression to these data. What is that problem? (*Hint:* If you're stuck, first draw the theoretical diagram you'd expect for P_i as a function of L_i, then plot Equation 1.25 onto the same diagram.)
   d. Suppose someone else took the data from the article and estimated:

$$P_i = 83.6 - 4.1L_i + e_i$$

   Is this the same result as that in Equation 1.25? If so, what definition do you need to use to convert this equation back to Equation 1.25?

7. Return to the housing price model of Section 1.5 and consider the following equation:

$$\hat{S}_i = 72.2 + 5.77P_i \qquad (1.26)$$

   where: S_i = the size (in square feet) of the *i*th house
   P_i = the price (in thousands of $) of that house

   a. Carefully explain the meaning of each of the estimated regression coefficients.
   b. Suppose you're told that this equation explains a significant portion (more than 80 percent) of the variation in the size of a house. Have we shown that high housing prices cause houses to be large? If not, what have we shown?
   c. What do you think would happen to the estimated coefficients of

13. Jaime Diaz, "Perils of Putting," *Sports Illustrated*, April 3, 1989, pp. 76–79.

this equation if we had measured the price variable in dollars instead of in thousands of dollars? Be specific.

8. If an equation has more than one independent variable, we have to be careful when we interpret the regression coefficients of that equation. Think, for example, about how you might build an equation to explain the amount of money that different states spend per pupil on public education. The more income a state has, the more they probably spend on public schools, but the faster enrollment is growing, the less there would be to spend on each pupil. Thus, a reasonable equation for per pupil spending would include at least two variables: income and enrollment growth:

$$S_i = \beta_0 + \beta_1 Y_i + \beta_2 G_i + \epsilon_i \qquad (1.27)$$

where: S_i = educational dollars spent per public school student in the *i*th state
Y_i = per capita income in the *i*th state
G_i = the percent growth of public school enrollment in the *i*th state

a. State the economic meaning of the coefficients of Y and G. (*Hint:* Remember to hold the impact of the other variable constant.)
b. If we were to estimate Equation 1.27, what signs would you expect the coefficients of Y and G to have? Why?
c. In 1995 Fabio Silva and Jon Sonstelie estimated a cross-sectional model of per student spending by state that is very similar to Equation 1.27.[14]

$$\hat{S}_i = -183 + 0.1422 Y_i - 5926 G_i \qquad (1.28)$$
$$n = 49$$

Do these estimated coefficients correspond to your expectations? Explain Equation 1.28 in common sense terms.
d. The authors measured G as a decimal, so if a state had a 10 percent growth in enrollment, then G equaled .10. What would Equation 1.28 have looked like if the authors had measured G in percentage points, so that if a state had 10 percent growth, then G would have equaled 10? (*Hint:* Write out the actual numbers for the estimated coefficients.)

14. Fabio Silva and Jon Sonstelie, "Did Serrano Cause a Decline in School Spending?" *National Tax Review*, June 1995, pp. 199–215. The authors also included the tax price for spending per

9. Your friend estimates a simple equation of bond prices in different years as a function of the interest rate that year (for equal levels of risk) and obtains:

$$\hat{Y}_i = 101.40 - 4.78X_i$$

where: Y_i = U.S. government bond prices (per \$100 bond) in the ith year
X_i = the federal funds rate (percent) in the ith year

a. Carefully explain the meanings of the two estimated coefficients. Are the estimated signs what you would have expected?
b. Why is the left-hand variable in your friend's equation $\hat{Y}$ and not Y?
c. Didn't your friend forget the stochastic error term in the estimated equation?
d. What is the economic meaning of this equation? What criticisms would you have of this model? (*Hint:* The federal funds rate is a rate that applies to overnight holdings in banks.)

10. Housing price models can be estimated with time-series as well as cross-sectional data. If you study aggregate time-series housing prices (see Table 1.2 for data and sources), you have:

$$\hat{P}_t = f(\overset{+}{GDP}) = 7404.6 + 19.8Y_t$$
$$n = 31 \text{ (annual 1964–1994)}$$

where: P_t = the nominal median price of new single-family houses in the U.S. in year t
Y_t = the U.S. GDP in year t (billions of current \$)

a. Carefully interpret the economic meaning of the estimated coefficients.
b. What is Y_t doing on the right side of the equation? Shouldn't it be on the left side?
c. Both the price and GDP variables are measured in nominal (or current, as opposed to real, or inflation-adjusted) dollars. Thus a major portion of the excellent explanatory power of this equation (more than 99 percent of the variation in P_t can be explained by Y_t alone) comes from capturing the huge amount of inflation that took place between 1964 and 1994. What could you do to eliminate the impact of inflation in this equation?
d. GDP is included in the equation to measure more than just inflation. What factors in housing prices other than inflation does the

TABLE 1.2 DATA FOR THE TIME-SERIES MODEL OF HOUSING PRICES

t	Year	Price(P_t)	GDP(Y_t)
1	1964	18,900	648.0
2	1965	20,000	702.7
3	1966	21,400	769.8
4	1967	22,700	814.3
5	1968	24,700	889.3
6	1969	25,600	959.5
7	1970	23,400	1010.7
8	1971	25,200	1097.2
9	1972	27,600	1207.0
10	1973	32,500	1349.6
11	1974	35,900	1458.6
12	1975	39,300	1585.9
13	1976	44,200	1768.4
14	1977	48,800	1974.1
15	1978	55,700	2232.7
16	1979	62,900	2488.6
17	1980	64,600	2708.0
18	1981	68,900	3030.6
19	1982	69,300	3149.6
20	1983	75,300	3405.0
21	1984	79,900	3777.2
22	1985	84,300	4038.7
23	1986	92,000	4268.6
24	1987	104,500	4539.9
25	1988	112,500	4900.4
26	1989	120,000	5250.8
27	1990	122,900	5546.1
28	1991	120,000	5724.8
29	1992	121,500	6020.2
30	1993	126,500	6343.3
31	1994	130,000	6736.9

P_t = the nominal median price of new single family houses in the U.S. in year t.
(Source: *The Statistical Abstract of the U.S.*)

Y_t = the U.S. GDP in year t (billions of current dollars).
(Source: *The Economic Report of the President*)

Note: EViews filename = HOUSE1

GDP variable help capture? Can you think of a variable that might do a better job?

11. The distinction between the stochastic error term and the residual is one of the most difficult concepts to master in this chapter.
    a. List at least three differences between the error term and the residual.

b. Usually, we can never observe the error term, but we can get around this difficulty if we assume values for the true coefficients. Calculate values of the error term and residual for each of the following six observations given that the true β_0 equals 0.0, the true β_1 equals 1.5, and the estimated regression equation is $\hat{Y}_i = 0.48 + 1.32X_i$:

Y_i	2	6	3	8	5	4
X_i	1	4	2	5	3	4

(*Hint:* To answer this question, you'll have to solve Equation 1.13 for ϵ and substitute Equation 1.15 into Equation 1.16.)
Note: filename = EX1

12. Look over the following equations and decide whether they are linear in the variables, linear in the coefficients, both, or neither.

a. $Y_i = \beta_0 + \beta_1 X_i^3 + \epsilon_i$

b. $Y_i = \beta_0 + \beta_1 \log X_i + \epsilon_i$

c. $\log Y_i = \beta_0 + \beta_1 \log X_i + \epsilon_i$

d. $Y_i = \beta_0 + \beta_1 X_i^{\beta_2} + \epsilon_i$

e. $Y_i^{\beta_0} = \beta_1 + \beta_2 X_i^2 + \epsilon_i$

13. What's the relationship between the unemployment rate and the amount of help-wanted advertising in an economy? In theory, the higher the unemployment rate, the lower the number of help-wanted ads, but is that what happens in the real world? Damodar Gujarati[15] tested this theory using time-series data for six years. You'd think that six years' worth of data would produce just six observations, far too few with which to run a reliable regression. However, Gujarati used one observation per quarter, referred to as "quarterly data," giving him a total of 24 observations. If we take his data set and run a linear-in-the-variables regression, we obtain:

$$\widehat{HWI}_t = 364 - 46.4UR_t \qquad (1.29)$$
$$n = 24 \text{ (quarterly 1962–1967)}$$

where: HWI_t = the U.S. help-wanted advertising index in quarter t
UR_t = the U.S. unemployment rate in quarter t

a. What sign did you expect for the coefficient of UR? (*Hint:* HWI rises as the amount of help-wanted advertising rises.) Explain your reasoning. Do the regression results support that expectation?

15. Damodar Gujarati, "The Relation Between the Help-Wanted Index and the Unemployment Index," *Quarterly Review of Economics and Business*, Winter 1968, pp. 67–73.

b. This regression is linear both in the coefficients and in the variables. Think through the underlying theory involved here. Does the theory support such a linear-in-the-variables model? Why or why not?
c. The model includes only one independent variable. Does it make sense to model the help-wanted index as a function of just one variable? Can you think of any other variables that might be important?
d. (optional) We have included Gujarati's data set, in Table 1.3 on our website, and on the EViews CD (as file HELP1). Use the EViews program (or any other regression software) to estimate Equation 1.29 on your own computer. Compare your results with Equation 1.29; are they the same?

TABLE 1.3

Observation	Quarter	HWI	UR
1	1962:1	104.66	5.63
2	1962:2	103.53	5.46
3	1962:3	97.30	5.63
4	1962:4	95.96	5.60
5	1963:1	98.83	5.83
6	1963:2	97.23	5.76
7	1963:3	99.06	5.56
8	1963:4	113.66	5.63
9	1964:1	117.00	5.46
10	1964:2	119.66	5.26
11	1964:3	124.33	5.06
12	1964:4	133.00	5.06
13	1965:1	143.33	4.83
14	1965:2	144.66	4.73
15	1965:3	152.33	4.46
16	1965:4	178.33	4.20
17	1966:1	192.00	3.83
18	1966:2	186.00	3.90
19	1966:3	188.00	3.86
20	1966:4	193.33	3.70
21	1967:1	187.66	3.63
22	1967:2	175.33	3.83
23	1967:3	178.00	3.93
24	1967:4	187.66	3.96

Note: filename = HELP1