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 are better off not watching in the making: sausages and econometric estimates."[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

1. Ed Leamer, "Let's take the Con out of Econometrics," *American Economic Review*, Vol. 73, No. 1, p. 37.

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

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 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 which part 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.

Uses of Econometrics

Econometrics has three major uses:

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

The simplest use of econometrics is description. We can use econometrics to quantify economic activity and measure marginal effects because econometrics allows us to estimate numbers and put them in equations that previously contained only abstract symbols. For example, consumer demand for a particular product often can be thought of as a relationship between the quantity demanded (Q) and the product's price (P), the price of a substitute (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 product. Econometrics actually allows us to estimate that

relationship based upon past consumption, income, and prices. In other words, a general and purely theoretical functional relationship like:

$$Q = \beta_0 + \beta_1 P + \beta_2 P_S + \beta_1 Yd \qquad (1.1)$$

can become explicit:

$$Q = 27.7 - 0.11P + 0.03P_S + 0.23Yd \qquad (1.2)$$

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 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 convince us that the true coefficient is indeed positive.

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 interested in this use of

3. It's of course naïve to build a model of sales (demand) without taking supply into consideration. Unfortunately, it's very difficult to learn how to estimate a system of simultaneous equations until you've learned how to estimate a single equation. As a result, we will postpone our discussion of the econometrics of simultaneous equations until Chapter 14. Until then, you should be aware that we sometimes will encounter right-hand-side variables that are not truly "independent" from a theoretical point of view.

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.

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. For example, economics typically is an observational discipline rather than an experimental one. "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. 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 used in nonexperimental 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

The specifications used in step 1 and the techniques used in step 3 differ widely between and within disciplines. Choosing the best specification 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 produce 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.

4. Clive Granger, "A Review of Some Recent Textbooks of Econometrics," *Journal of Economic Literature*, Vol. 32, No. 1, p. 117.

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 always should 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 iPhones demanded will increase if the price of those phones decreases (holding everything else constant), but not many people can put specific numbers into an equation and estimate *by how many* iPhones 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.

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 one or more equations. For example, in Equation 1.1:

$$Q = \beta_0 + \beta_1 P + \beta_2 P_S + \beta_1 Yd \qquad (1.1)$$

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 equations. 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 increase 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 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 often is 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.

Single-Equation Linear Models

The simplest single-equation regression model is:

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

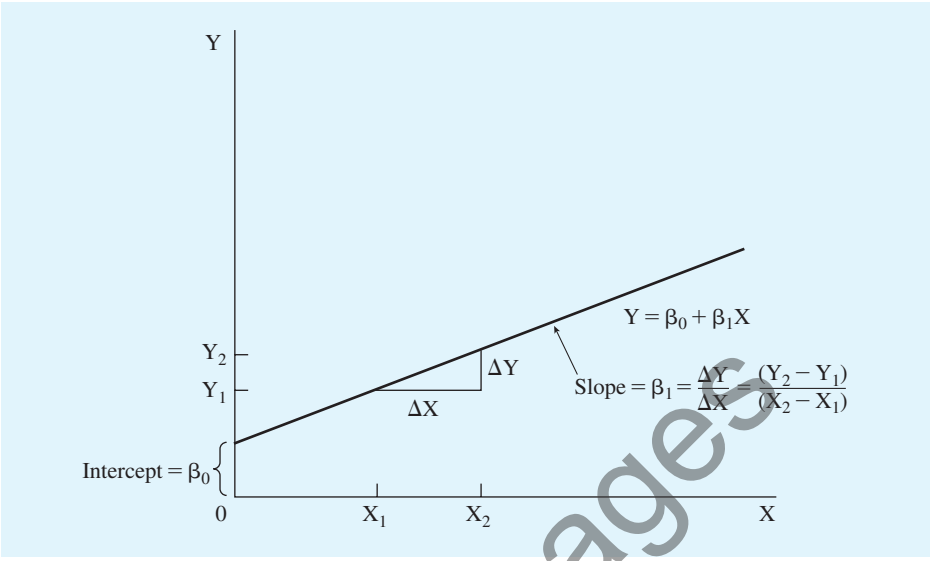

Figure 1.1 Graphical Representation of the Coefficients of the Regression Line

The graph of the equation $Y = \beta_0 + \beta_1 X$ is linear with a constant slope equal to $\beta_1 = \Delta Y / \Delta X$.

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 it's the only equation specified. The model is linear because if you were to plot Equation 1.3 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 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 one-unit increase in X. 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 by one 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.

If linear regression techniques are going to be applied to an equation, that equation *must* be linear. An equation is **linear** if plotting the function in terms of X and Y generates a straight line; for example, Equation 1.3 is linear.[5]

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

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

5. Technically, as you will learn in Chapter 7, this equation is linear in the coefficients β_0 and β_1 and linear in the variables Y and X. 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.

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

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

Equation 1.4 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 13-year-old girls is 5 feet, then 5 feet is the expected value of a girl's height given that she is 13. The deterministic part of the equation may be written:

$$E(Y|X) = \beta_0 + \beta_1 X \tag{1.5}$$

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).

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 13-year-old girls are 5 feet tall. As a result, the stochastic element (ϵ) must be added to the equation:

$$Y = E(Y|X) + \epsilon = \beta_0 + \beta_1 X + \epsilon \tag{1.6}$$

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 the dependent variable.
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.
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

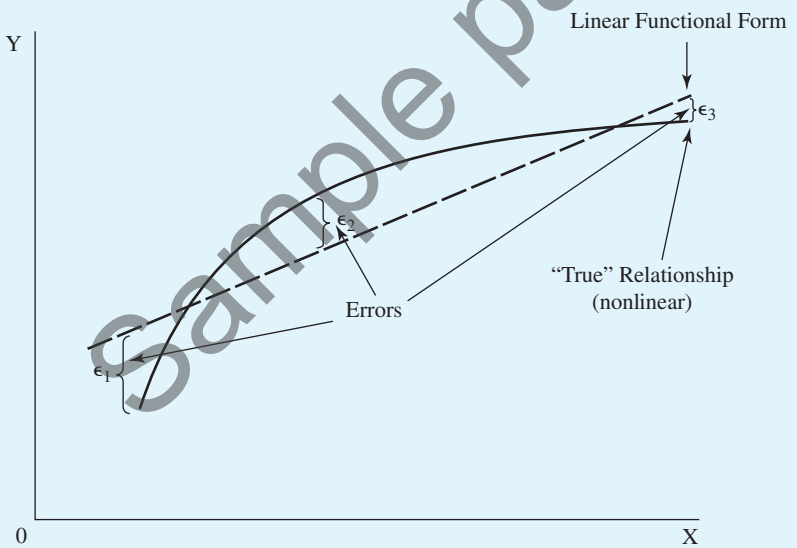

Figure 1.2 Errors Caused by Using a Linear Functional Form to Model a Nonlinear Relationship

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.

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.

Extending the Notation

Our regression notation needs to be extended to allow the possibility of more than one independent variable and to include reference to the number of observations. A typical observation (or unit of analysis) is an individual person, year, or country. For example, a series of annual observations starting in 1985 would have $Y_1 = Y$ for 1985, Y_2 for 1986, etc. If we include a specific reference to the observations, the single-equation linear regression model may be written as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (i = 1, 2, \ldots, N) \tag{1.7}$$

where:
Y_i = the ith observation 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

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 ith observation of the first independent variable
X_{2i} = the ith observation of the second independent variable
X_{3i} = the ith observation of the third independent variable

then all three variables can be expressed as determinants of Y.

The resulting equation is called a **multivariate** (more than one independent variable) linear **regression model:**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \tag{1.8}$$

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* 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 very difficult to run controlled economic experiments,[7] 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.

7. Such experiments are difficult but not impossible. See Section 16.1.

This material is pretty abstract, so let's look at two examples. As a first example, consider an equation with only one independent variable, a model of a person's weight as a function of their height. The theory behind this equation is that, other things being equal, the taller a person is the more they tend to weigh.

The dependent variable in such an equation would be the weight of the person, while the independent variable would be that person's height:

$$\text{Weight}_i = \beta_0 + \beta_1 \text{Height}_i + \epsilon_i \tag{1.9}$$

What exactly do the "i" subscripts mean in Equation 1.9? Each value of i refers to a different person in the sample, so another way to think about the subscripts is that:

$$\begin{aligned}
\text{Weight}_{\text{woody}} &= \beta_0 + \beta_1 \text{Height}_{\text{woody}} + \epsilon_{\text{woody}} \\
\text{Weight}_{\text{lesley}} &= \beta_0 + \beta_1 \text{Height}_{\text{lesley}} + \epsilon_{\text{lesley}} \\
\text{Weight}_{\text{bruce}} &= \beta_0 + \beta_1 \text{Height}_{\text{bruce}} + \epsilon_{\text{bruce}} \\
\text{Weight}_{\text{mary}} &= \beta_0 + \beta_1 \text{Height}_{\text{mary}} + \epsilon_{\text{mary}}
\end{aligned}$$

Take a look at these equations. Each person (observation) in the sample has their own individual weight and height; that makes sense. But why does each person have their own value for ϵ, the stochastic error term? The answer is that random events (like those expressed by ϵ) impact people differently, so each person needs to have their own value of ϵ in order to reflect these differences. In contrast, note that the subscripts of the regression coefficients (the βs) don't change from person to person but instead apply to the entire sample. We'll learn more about this equation in Section 1.4.

As a second example, let's look at an equation with more than one independent variable. Suppose we want to understand how wages are determined in a particular field, perhaps because we think that there might be discrimination in that field. The wage of a worker would be the dependent variable (WAGE), but what would be good independent variables? What variables would influence a person's wage in a given field? Well, there are literally dozens of reasonable possibilities, but three of the most common are the work experience (EXP), education (EDU), and gender (GEND) of the worker, so let's use these. To create a regression equation with these variables, we'd redefine the variables in Equation 1.8 to meet our definitions:

Y = WAGE = the wage of the worker
X_1 = EXP = the years of work experience of the worker
X_2 = EDU = the years of education beyond high school of the worker
X_3 = GEND = the gender of the worker (1 = male and 0 = female)

The last variable, GEND, is unusual in that it can take on only two values, 0 and 1; this kind of variable is called a dummy variable, and it's extremely useful when we want to quantify a concept that is inherently qualitative (like gender). We'll discuss dummy variables in more depth in Sections 3.3 and 7.4.

If we substitute these definitions into Equation 1.8, we get:

$$WAGE_i = \beta_0 + \beta_1 EXP_i + \beta_2 EDU_i + \beta_3 GEND_i + \epsilon_i \qquad (1.10)$$

Equation 1.10 specifies that a worker's wage is a function of the experience, education, and gender of that worker. In such an equation, what would the meaning of β_1 be? Some readers will guess that β_1 measures the amount by which the average wage increases for an additional year of experience, but such a guess would miss the fact that there are two other independent variables in the equation that also explain wages. The correct answer is that β_1 gives us the impact on wages of a one-year increase in experience, *holding constant* education and gender. This is a significant difference, because it allows researchers to control for specific complicating factors without running controlled experiments.

Before we conclude this section, it's worth noting that the general multivariate regression model with K independent variables is written as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i \qquad (1.11)$$

where i goes from 1 to N and indicates the observation number.

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.[8]

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 data for actual Xs and Ys. Although the theoretical equation is purely abstract in nature:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (1.12)$$

8. The order of the subscripts doesn't matter as long as the appropriate definitions are presented. We prefer to list the variable number first (X_{1i}) because we think it's easier for a beginning econometrician to understand. However, as the reader moves on to matrix algebra and computer spreadsheets, it will become common to list the observation number first, as in X_{i1}. Often the observational subscript is deleted, and the reader is expected to understand that the equation holds for each observation in the sample.

the estimated regression equation has actual numbers in it:

$$\hat{Y}_i = 103.40 + 6.38X_i \tag{1.13}$$

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 1.12 have been replaced with *estimates* of those coefficients like 103.40 and 6.38 in Equation 1.13. We can't actually observe the values of the true[9] regression coefficients, so instead we calculate estimates of those coefficients from the data. The estimated regression coefficients, 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

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{1.14}$$

is the empirical counterpart of the theoretical regression Equation 1.12. The calculated estimates in Equation 1.13 are examples of the 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 prediction of $E(Y_i|X_i)$ from the regression equation. The closer these $\hat{Y}$s are to the Ys in the sample, 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):

$$e_i = Y_i - \hat{Y}_i \tag{1.15}$$

9. 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.