CHAPTER

1 Introduction to Econometrics

It is interesting that people try to find meaningful patterns in things that are essentially random.

—Data, Star Trek

LEARNING OBJECTIVES

Upon completing the work in this chapter, you will be able to:

- Define and describe the basics of econometrics
- Describe how to do an econometric study

Jaime Escalante was born in Bolivia in 1930. He immigrated to the United States in the 1960s, hoping for a better life. After teaching himself English and working his way through college, he became a teacher at Garfield High School in East Los Angeles. Jaime believed strongly that higher math was crucial for building a successful career, but most of the students at Garfield High, many of whom came from poor backgrounds, had very weak math skills. He worked tirelessly to transform these kids into math whizzes. Incredibly, more than a quarter of all the Mexican-American high school students who passed the AP calculus test in 1987 were taught by Jaime.

Hollywood made a movie of Jaime's story called "Stand and Deliver." If you haven't seen that movie, you've probably seen one of the other dozens with a similar plot. An inspiring and unconventional teacher gets thrown into an unfamiliar environment filled with struggling or troubled kids. The teacher figures out how to reach the kids, they perform well in school, and their lives change forever.

We all have stories of an inspiring teacher we once had. Or a terrible teacher we once had. Meanwhile, school boards everywhere struggle with the question of how to teach kids and turn them into economically productive adults. Do good teachers really make all the difference in our lives? Or do they merely leave us with happy

memories? Not every school can have a Jaime Escalante. Is more funding for public schools the answer? Smaller class sizes? Better incentives for teachers? Technology?

Econometrics can provide answers to big questions like these.

WHAT IS ECONOMETRICS?

Humans have been trying to make sense of the world around them for as long as anyone knows. Data bombard our senses: movements in the night sky, the weather, migrations of prey, growth of crops, spread of pestilence. We have evolved to have an innate curiosity about these things, to seek patterns in the chaos (empirics), then explanations for the patterns (theories). Much of what we see around us *is* random, but some of it is not. Sometimes our lives have depended on getting this right: predicting where to find fish in the sea (and being smart enough to get off the sea when a brisk nor'easter wind starts to blow), figuring out the best time to plant a crop, or intervening to arrest the spread of a plague. A more complex world gives us ever more data we have to make sense of, from climate change to Google searches to the ups and downs of the economy.

Econometrics is about making sense of economic data (literally, it means "economy measurement"). Often, it is defined as the application of statistics to economic data, but it is more than that. To make sense of economic data, we usually need to understand something about the unseen processes that create these data. For example, we see differences in people's earnings and education (years of completed schooling). Econometric studies consistently find that there is a positive relationship between the two variables. Can we use people's schooling to predict their earnings? And if we *increase* people's schooling, can we say that their earnings will increase?

These are two different questions, and they get at the hardest part of econometrics—distilling causation from correlation. We may use an econometric model to learn that people with a college degree earn more than those without one. That is a predictive, or correlative, relationship. We don't know whether college graduates earn more because of useful things they learned in college—that is, whether college *causes* higher earnings. College graduates tend to have high IQ, and they might have earned a lot regardless of whether or not they went to college. Mark Twain (who was not educated beyond elementary school) once said: "I've never let my school interfere with my education." He might have had a point.

Often, an econometrician's goal is to determine whether some variable, X, causes an outcome, Y. But not all of econometrics is about causation. Sometimes we want to generate predictions and other times test a theory. Clearly defining the purpose of an econometrics research project is the first step

toward getting credible and useful results. The second step is to formulate your research design and specify your econometric model, and the final step is to apply statistical theory to answer the question posed in step 1.

Most of your first econometrics course focuses on step 3, but don't forget steps 1 and 2! Throughout the book, we will remind you of these steps. Next, we discuss each of the three steps to put the rest of the book in context.

STEP 1: WHAT DO YOU WANT TO DO?

The first step in doing econometrics is to define the purpose of the modeling. It is easy to skip this step, but doing so means your analysis is unlikely to be useful.

Your purpose should be concrete and concise. "I want to build a model of the economy" is not enough. What part of the economy? What do you want to learn from such a model? Often, if you can state your purpose in the form of a question, you will see whether you have defined it adequately.

Here are some examples.

Do Good Teachers Produce Better Student Outcomes?

To estimate whether good teachers improve life outcomes, we first need to measure teacher quality. In a 2014 study, Raj Chetty, John Friedman, and Jonah Rockoff constructed measures of how much an above-average teacher improves students' test scores over what they would have been with an average teacher. These are called "value-added" (VA) measures of teacher quality and were estimated using detailed data on elementary school records from a large urban school district. This research was deemed so important that it was presented in not one but two papers in the most prestigious journal in economics, the *American Economic Review*.[1]

Chetty and his coauthors used econometrics with their VA measures to show that replacing an average teacher with a teacher whose VA is in the top 5% would increase students' earnings later in life by 2.8%. This might seem small, but the average 12-year-old in the United States can expect lifetime earnings of $522,000,[2] so a 2.8% earnings bump is worth about $14,500 per student. Multiply that by 20 kids per classroom and an excellent teacher starts to look really valuable. It works the other way too—teachers with low VA potentially have large negative effects on lifetime earnings.

Does the Law of Demand Hold for Electricity?

In microeconomic theory, the law of demand predicts that when the price of a good rises, demand for the good falls. Does this theoretical prediction

hold up in the real world? Is the own-price elasticity of demand really negative? How large is it? Finding a negative correlation between price and demand is consistent with economic theory; finding the opposite is not.

Katrina Jessoe and David Rapson asked this question using data on residential electricity consumption.[3] They conducted an experiment in which they divided homes randomly into three groups. The first group faced electricity prices that jumped by 200–600% on certain days of the year. The second group faced the same price rises but also were given an electronic device that told them in real time how much electricity they were using. The third group was the control group: they experienced no change in their electricity prices.

Jessoe and Rapson used econometrics to estimate that consumers in the first group did not change their consumption significantly compared with the control group—they had a price elasticity of demand close to zero. However, the second group had a price elasticity of demand of –0.14. Conclusion: the law of demand holds for electricity, but only if consumers know how much electricity they are using in real time. Without this knowledge, they don't know how much electricity is used when they run the air conditioner or switch on a light, so they don't respond to a price change.

Is It Possible to Forecast Stock Returns?

Lots of people think they can make money in the stock market. We often receive emails informing us of the next greatest stock tip. Business TV channels are full of people yelling about how to make money in the stock market. Every time the market crashes, there's a great story about the genius investor who saw it all coming and made money during the crash.[4] But if it's so easy to make money in the stock market, why isn't everyone doing it?

Based on the theory of efficient financial markets, many economists cast a skeptical eye on claims that the stock market is highly predictable. If everyone knew the market was going to go up, then it would have already done so. However, economic theory also predicts that financial investments should have returns in proportion to risk. Riskier investments should have higher returns on average. So, if you can measure risk in the stock market, then you should be able to predict returns to some extent.

Ivo Welch and Amit Goyal took a large number of variables that people claimed could predict stock returns and used econometrics to test whether any of them actually could.[5] They conclude that there is little, if any, statistical evidence of stock return predictability. So next time you hear a prognosticator claiming that the stock market is about to crash because it crashed the last seven times the president went skiing on a Tuesday . . . or something . . .., change the channel!

STEP 2: FORMULATE YOUR RESEARCH DESIGN AND SPECIFY THE ECONOMETRIC MODEL

This step typically requires some economic theory, common sense, and a little cleverness. It is where you take your abstract objective from step 1 and convert it into an econometric model with data that can answer your questions.

Making a good choice about what data to collect and use determines whether you will be able to meet your objective. Let's look again at the three studies we highlighted above.

Do Good Teachers Produce Better Student Outcomes?

Microeconomic theory of the firm provides us with a theory of how a teacher might affect earnings. It's called human capital theory.[6] Human capital theory predicts that workers are paid the marginal value product of their labor (MVPL). A firm will not hire a worker unless the additional value she produces (her MVPL) is at least as large as what the firm will have to pay the worker (i.e., the wage). Characteristics that raise workers' productivity, like intelligence, ability to concentrate and willingness to work hard, should be associated with higher wages. Having had a good teacher is one characteristic that may raise productivity.

One possible research design would be to build an econometric model of the determinants of test scores. If students in teacher A's class get better than average test scores, then teacher A must be a good teacher. There is a big problem with that approach. Classes with a lot of students from disadvantaged backgrounds will tend to get lower-than-average scores no matter how good the teacher is (unless the teacher is the subject of a Hollywood movie).

This is why Chetty and his coauthors developed their VA measures.[7] Their method first predicts test scores for thousands of students based on variables such as last year's test score and family socioeconomic characteristics. Next, they look at how well each student does relative to the prediction. If the students in a class tend to do better than predicted, then Chetty and coauthors assign a high VA score to the teacher. They conduct a series of tests of their VA measure. For example, they look at what happens when an average teacher leaves a school and is replaced by a teacher with a higher VA. They find that test scores jump up from the previous year, which validates their method.

Does the Law of Demand Hold for Electricity?

Microeconomic theory posits that the price and quantity in a market are determined at the point where supply equals demand. When some exogenous shock (like a new invention) shifts the supply outward, the price drops to convince

consumers to buy more. When demand increases, the price rises to convince suppliers to produce more. When testing the law of demand, an econometrician wants to see how much consumers respond when a supplier changes the price. The problem is that often in the real world the price is high because consumers really like the product.

When the weather gets hot, consumers turn on their air conditioners and their electricity consumption goes up. This weather-induced increase in demand could cause electricity prices to go up.[8] It would generate a positive correlation between consumption and price. Should we conclude, then, that the law of demand is false? Of course not, because the positive correlation comes from high demand causing high prices, not high prices causing more demand.

One way to think about whether you have a good research design is to imagine what experiment you would run if you could. Jessoe and Rapson went one better and actually ran it. They convinced an electric utility to let them raise prices for a random set of customers on hot days and keep prices the same for other customers. Because they were controlling prices themselves, and because they randomly assigned who got the high prices rather than cherry picking receptive customers, they could be confident they were really measuring consumer responses. This is an example of experimental economics.

Even if you can't run the experiment, thinking about how you would conduct that experiment can help you figure out whether you have data that can answer your research question.

Is It Possible to Forecast Stock Returns?

Finance theory says that stock returns should be higher when investors are more risk averse. (A higher return is needed to incentivize these investors to take on more risk and invest.) Researchers have proposed many measures of risk – often referred to as systematic risk factors. For example, if you have a high chance of losing your job, you are likely to be more risk averse than otherwise. You would only want to put your money in something risky like stocks if the price were low enough and the future expected gains high enough to make it worth the risk.

Welch and Goyal used this theory and the research from past studies to come up with a list of predictors to test. Without theoretical guidance on which predictors to consider, the possibilities would be endless.

Even with a theoretically motivated group of predictors, demonstrating whether a predictor works is hard because we don't *really* know how the predictor was chosen. A previous researcher may have engaged in data snooping, which means they searched repeatedly until they found a variable that correlated significantly with stock returns and then made up an economic story about why it measures risk aversion. If you search hard enough, you will find a

significant looking but meaningless correlation. Welch and Goyal account for this possibility using what is called an *out-of-sample test*. They fit econometric models using data prior to 1965, and then they see whether the predictors that perform well prior to 1965 continue to work after 1965.

In each of these studies, the researchers applied some economic theory, some common sense, and a little cleverness. The best sources of theoretical insights are the intermediate theory courses that most likely were a prerequisite for this econometrics course. Throughout this book, we will refer to what we learned in our theory courses as a useful resource in building econometric models.

The researchers each thought about mathematical form of their models. Does Y increase linearly with X? Is this relationship likely to be quadratic instead of linear? Logarithmic? Which control variables should be included in the equation? We can use econometrics to test whether the relationship between X and Y is linear or nonlinear, and what mathematical form is best to predict an outcome of interest. For example, we might find that there is a significant relationship between X and Y that is evident using a nonlinear model but not a linear one. In short, we need both economic theory and mathematics—plus a little experimentation with functional forms—to come up with the model we want to estimate.

STEP 3: APPLY STATISTICAL THEORY

Statistical theory helps us fill the gap between the numbers we compute from our data and the broader world. There is a gap because no econometric model can perfectly predict every data point and because we usually only observe data from a sample of the population.

Suppose you are trying to predict earnings of individuals for a population of 100,000 people. Now, imagine you get really lucky: someone hands you data on earnings and years of education for the *whole* population. That's right, 100,000 people. (It sounds like a lot of data, but Aaron and Ed work with *samples* a lot larger than this sometimes, and it's puny compared to what Google works with!) We said "really lucky" because we almost never have data on whole populations.

With data on the whole population, you could fit a regression model using the methods we'll learn in Chapter 2. This model would produce the predicted income for each individual based on his or her education level. Suppose you find, with this population model, that the expected earnings for a person with 16 years of schooling is $50,000.

But not every person with 16 years of education will earn exactly $50,000. Some will earn more, and others will earn less. We cannot expect, using only information on education, to predict everyone's income perfectly. The model has an error. To be precise, let's call this the population error.

Perfect prediction is usually not a realistic goal for our analysis. So it's OK that the model has an error. However, the error creates another problem. In the real world, we typically observe only a sample of the population. You might only have 1,000 observations with which to make predictions about a population of 100,000. You know your sample model will have an error, but how much of the error comes from differences between the sample and population and how much is due to the population error?

This is where statistics comes in. Statistics is the science of using sample data to make statements about a whole population.

How do we do this? The first thing you should worry about is how the sample of 1,000 got chosen from the population of 100,000. To be able to say something about the whole population, the sample you used to estimate the model has to be representative of the whole population. If your sample is not representative of the population, the errors in your predictions will reflect that lack of representativeness. "Representative" means "as if drawn randomly from the population." For example, if your sample includes mostly low-ability people, it is not representative of a population that includes both low- and high-ability individuals. The representativeness of the sample is so important that we devote a chapter to it later in this book (Chapter 10).

Given a representative sample, we need to make some assumptions about what the population errors look like in order to use our sample data to make statements of probability about the whole population. Once we learn the basic econometric regression model, most of the rest of this book will be dedicated to testing whether those assumptions about the population error are correct, and what to do about it if they are not.

What generates errors in the population model? The answer is that anything that explains earnings and is not in the model gets reflected in the error. That "anything" could include variables that we think should be in the model, because theory, experience, or intuition tell us so, but for which we do not have information. The effects of innate intelligence, drive, willingness to take risks, and other variables that might systematically influence earnings fall into the error if we do not include them explicitly in our model.

Missing and unobserved variables, we shall see, can cause problems, particularly when we want to estimate a *causal* relationship between X and Y. That is what Chapters 11 and 12 are about. There we will see that if X is exogenous, like if education levels were assigned randomly at birth, then we can estimate the causal relationship between X and Y even if other relevant variables are not in the model. Often, though, the right-hand-side variables in our model are not exogenous, like when people decide how much education to get, and unseen variables like ability explain both education and earnings. Then it becomes difficult—sometimes impossibly so—to isolate the effect of the included variable (education) from the effect of the omitted one (ability).

Even when relevant variables are left out of the model, the model might be useful for predicting Y using information on X. Knowing how much education someone has allows us to get better predictions of someone's earnings. Including additional relevant variables in our model might improve our predictive power.

We cannot control for everything. There is a stochastic or random component of the error, which is often referred to as "white noise." Albert Einstein wrote: "God does not play dice with the universe." But when it comes to statistics, the relationship we see between X and Y has an error thrown in, like rolling dice.

AN ILLUSTRATION WITH THE POPULATION MEAN

Here's a simple illustration of the population error. Suppose we have reliable data on income (Y) for *everyone* in a population. We could use these data to calculate the population's true mean income, which we could designate as μ (the Greek letter "mu"). You know how to calculate it: take the sum of everyone's income and divide it by the size of the population.

We already pointed out that we almost never have data on the entire population—only a sample of it. Instead of gathering income data on everyone in the population (which may be prohibitively expensive and infeasible), we can collect data from a random sample of size N from the population. Assume the sampling is truly random, which makes the sample representative of the whole population. We can use these data to estimate the sample mean, using the estimator $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$. Unless we are impossibly lucky, this will not equal μ. What we get will depend on the "luck of the draw" that gave us the sample.

We could go out and survey a different random sample of the population, and we'd get a different estimate of μ. We could repeat this process another 9,998 times, and we'd have 10,000 different estimates of μ. Sometimes we will get an estimate that is high, and sometimes it will be low. Every now and then it will be *very* high or *very* low.

If we were to pick one of the 10,000 estimates at random, we would have no reason to expect it to be too high or too low. Formally, this means that the sample mean is unbiased as an estimate of the population mean. With this knowledge and a little more—including an estimate of the population variance—statistics lets us make statements of probability about the true population mean using data from a single sample. That, in a nutshell, is the power of statistics. In Chapters 4–6, we will learn how to make similar kinds of statements of probability using econometric models.

Econometrics is challenging because it integrates all three of these fields—economics, mathematics, and statistics—into one.

PUTTING IT TOGETHER: POVERTY AND TEST SCORES

We started this chapter with a story about how a good teacher can improve student test scores. Many other variables also affect test scores. Like poverty. Some schools are located in poor neighborhoods, while others are in rich ones. Children who grow up in poor households might not have the same opportunities to learn and expand their minds through travel, the internet, or other means, as children in rich households. They may be consumed with worry about where the next meal will come from rather than focusing on school. Schools in poor neighborhoods might have poorer facilities, fewer extracurricular activities, less access to advanced placement courses, and less overall funding per pupil than schools in rich neighborhoods.

How does poverty relate to student performance? To answer this question, we first need to measure "student performance" and "poverty."

The government of California measures school performance using a single number, the academic performance index (API), constructed from students' scores on statewide standardized tests. The API ranges from a low of 200 to a high of 1,000. We can use the API as our measure of average student performance at each California school.

Measuring poverty is more problematic. We do not know the household per capita incomes for each student at each school, so we cannot measure poverty directly. However, we do have an indirect measure or proxy for poverty. The US National School Lunch Program provides free or reduced-price lunches to schoolchildren from low-income households. We know the share of students eligible for free or reduced-price lunches at each school in California. We can use free-lunch eligibility (FLE) as an indicator of the share of students from households with incomes below the poverty line at each school.

Table 1.1 presents the API and FLE at 20 randomly chosen California schools in 2013—from a total population of 5,765 schools. Our first challenge will be to figure out the relationship between FLE and API, using these data. In the next chapter, we will learn how to use econometric tools to predict schools' API based on their FLE. In chapter 3, we will take this a step further by asking what other variables can be used to predict API and how to include these other variables in our econometric model. In Chapters 4 and 5, we will learn how to use findings from samples of schools like this one to make statements about FLE and API in *all* schools in California. Subsequent chapters will address the problems that frequently appear when using econometric findings from a sample of data to generalize to populations outside the sample. Most of these problems have to do with the error term.

We'll see that we can learn a great deal by using econometrics to estimate the correlations between variables like FLE and API. As we've already pointed out, making statements about causation is more complicated. In the present example, suppose we find that high FLE predicts low API. Does this mean that

Table 1.1 Academic Performance Index (API) and Free Lunch Eligibility (FLE) at 20 California Elementary Schools in 2013

School	County	API	FLE
Joe A. Gonsalves Elementary	Los Angeles	960	16
Old River Elementary	Kern	849	0
Sierra Vista Elementary	Kern	722	96
West Portal Elementary	San Francisco	914	44
Isabelle Jackson Elementary	Sacramento	754	83
Rio Vista Elementary	Orange	796	90
Poplar Avenue Elementary	Butte	802	80
Cloverly Elementary	Los Angeles	903	46
Creative Arts Charter	San Francisco	844	33
Carolyn A. Clark Elementary	Santa Clara	963	6
Raymond Elementary	Orange	824	69
Fernangeles Elementary	Los Angeles	730	100
Rainbow Ridge Elementary	Riverside	826	90
Cyrus J. Morris Elementary	Los Angeles	882	29
Benjamin Franklin Elementary	Riverside	882	36
Salvador Elementary	Napa	736	65
Bowers Elementary	Santa Clara	788	59
Vintage Parkway Elementary	Contra Costa	830	54
Balboa Magnet Elementary	Los Angeles	981	22
Selby Lane Elementary	San Mateo	730	80
	Mean	835.80	54.90
	Std. deviation	81.14	31.00

SOURCE: California Department of Education (http://www.cde.ca.gov/ta/ac/ap/apidatafiles.asp).

if the government increased funding for free school lunches by, say, 10%, API would decrease?

You're probably thinking, "Wait a minute, that's a different question!" If so, you're right. Our motivation for using FLE to predict API was that FLE reflects, or is a proxy for, poverty. If the state provided more funding for school lunches, there's no reason to think poverty would go up. On the contrary, more poor students might become well nourished, and their school performance might improve, so increased funding for free lunches could increase API! Clearly, correlation and causation are very different matters in this case. FLE might be a good predictor of (lower) API, but that doesn't mean that a government program that increases FLE will decrease API.

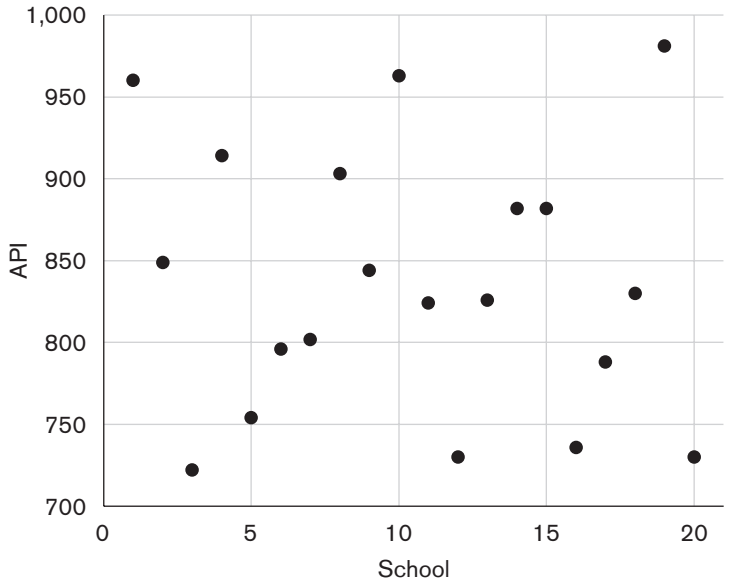

Figure 1.1 2013 API in 20 California elementary schools. Academic performance varies widely across these 20 schools.

We could think of other possible interventions to increase API. What if welfare payments to poor households with school-age children increased? Or the state provided new funding to decrease class sizes in schools? Would API increase? Addressing these sorts of cause-and-effect questions is a focus of what many econometricians do, so we include them in this book (Chapter 11)—after we master the classical econometric regression model.

FROM STATISTICS TO ECONOMETRICS

Before we get into the econometrics, let's get some mileage out of the statistics courses you've taken. We can plot student performance across schools (figure 1.1). In figure 1.1, each of the 20 schools is lined up, in no particular order, along the horizontal axis, and the schools' APIs are measured along the vertical axis.

This figure shows that academic performance varies widely across these 20 schools, from less than 750 to over 950. The average is 835.80, and the standard deviation is 81.14. (You can use a spreadsheet program like EXCEL to verify this.)

What if we use the average API to predict the API of any given school? That would mean using a model like this one, from your introductory statistics course, in which the API of school i is denoted by Y_i:

$$Y_i = \mu + \varepsilon_i.$$

In this statistical model, we imagine that the APIs we see for different schools are generated by taking the population mean API (which we call μ, a constant) and then throwing in an error, ε_i, which we hope is random. The sum of μ and ε_i gives us the observed school performance outcome for school i, Y_i.

If we used our sample mean to predict the API at Sierra Vista Elementary in Kern County, it would do a lousy job. This school has an API of 722, which is much lower than the mean of 835.80.

What is it about Sierra Vista Elementary, and many other schools in these data, that makes the mean for *all* schools a poor predictor of API for a *particular* school? If we could find some variable, X_i, that helped explain differences in APIs across schools, we could use it to construct a new (hopefully better) prediction model. In the new model, Y_i might be a simple linear function of X_i:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Notice what we've done here. We've replaced μ with $\mu_i = \beta_0 + \beta_1 X_i$. A school's expected API is no longer constant; it depends on X_i. X_i could be poverty. It could be that schools in poor neighborhoods perform more poorly, on average, than schools in rich neighborhoods. Sierra Vista Elementary is only one school among many in California, but it is striking that 96% of kids at this school receive free school lunches, compared to only 22% at the school with the highest API score in Table 1.1 (Balboa Magnet Elementary in Los Angeles; API = 981).

A Theoretical Foundation for an Econometric Model of Academic Performance

What kind of economic theory would predict an impact of poverty on API? We need a model of how schools achieve a high API. Call it an "API production function."

You studied production functions in microeconomic theory courses. Firms combine labor, capital, and other inputs to produce an output. The production function describes the technology that turns inputs into output. Do schools have a production function that turns inputs into API? What would those inputs be?

Classrooms, teachers, teaching aids, specialists, books, art supplies, computers, electricity to run everything—all of these are inputs in our API production function, and all of them cost money. Schools face budget constraints, and

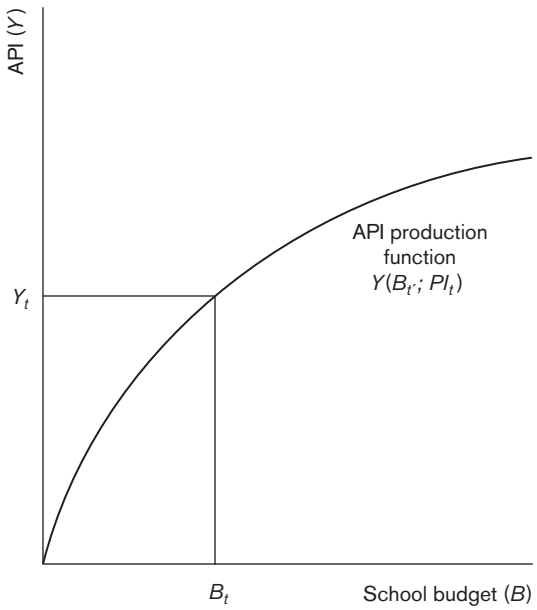

Figure 1.2 Schools convert budgets per pupil into academic performance, indexed by the API.

despite some states' best efforts to equalize spending per pupil, significant inequalities remain. Schools in high-income districts can pass bond measures and run successful fundraising campaigns to support extracurricular activities, build new facilities, and hire teaching specialists. Schools in poor districts have a tough time keeping up.

Schools with larger budgets per pupil (B) can buy more inputs, which in turn help them achieve a higher API, as shown in figure 1.2. This figure assumes that the impact of B on API is positive, but at high levels of per-pupil budget diminishing marginal returns eventually set in, like in a conventional production function.

The other critical input in the API production function, of course, is students. Students come from households, which face their own budget constraints. Households take their limited budgets (BH) and allocate them to goods and services that directly benefit their children ("child investments"), as well as to other goods and services. Nutritious foods, clothing, medical care, books, the internet, parents' time spent nurturing their children's minds, travel—all of these are examples of private investments that can prepare children to do well in school. Poor households face severe constraints on the investments they can make in their children. For a disproportionate share of children in poor US households, English is a second language. This makes the school's job more challenging. As household incomes rise, so do private investments per student (PI), as shown in figure 1.3.

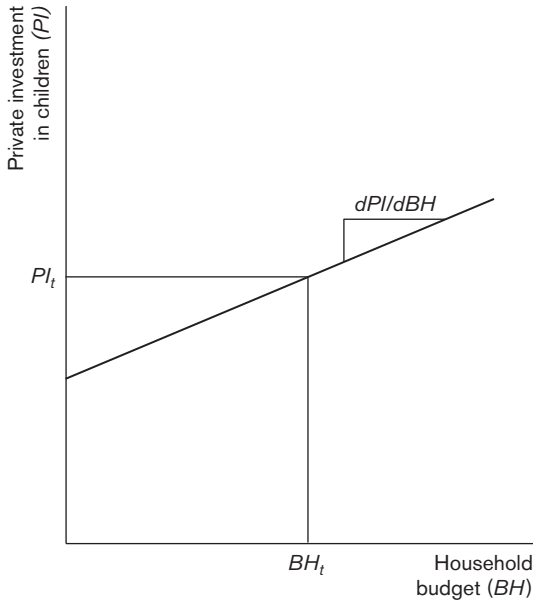

Figure 1.3 Private investments per student (PI) increase with household budgets (*BH*), assuming they are a normal good.

You will notice that we included PI in the API production function in figure 1.2. There is no axis in the diagram for PI because PI is an exogenous "shifter" in the API production function. Schools in neighborhoods where PI is high have a head start in turning *B* into API (upward shift in the API production function) compared with schools in poor neighborhoods, as illustrated in figure 1.4.

The model depicted in figures 1.2–1.4 provides a theoretical grounding for an econometric analysis of the relationship between poverty (proxied by FLE) and children's academic performance (measured by the API). Our theoretical model predicts that there is a negative relationship between poverty and academic performance: as school lunch eligibility increases, academic performance declines. This theory is particularly useful because it gives reasons why poverty might not only be *correlated* with lower school performance, but also actually *cause* it. It is also useful because it offers some insights into what other variables we might need to have in our model, besides poverty. We'll return to these two points in Chapters 2 and 3.

We can begin by plotting API against FLE, as in figure 1.5.

It looks like there's a negative relationship. The next step is to use econometrics to estimate it for these 20 schools. Let's see what we can learn from FLE to help us predict API. We'll begin by estimating a linear model in which FLE is the only variable explaining API, that is,

$$Y_i = b_0 + b_1 X_i + e_i,$$

where *Y* denotes API and *X* denotes FLE.

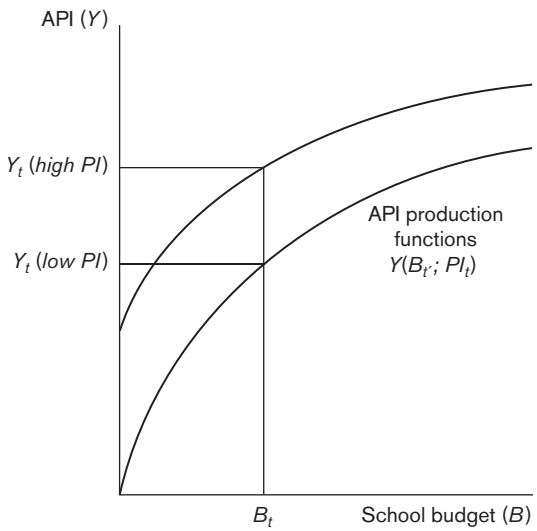

Figure 1.4 Private investments per pupil (PI) shift the API-production function upward; the same per-pupil school budget produces a higher average student performance.

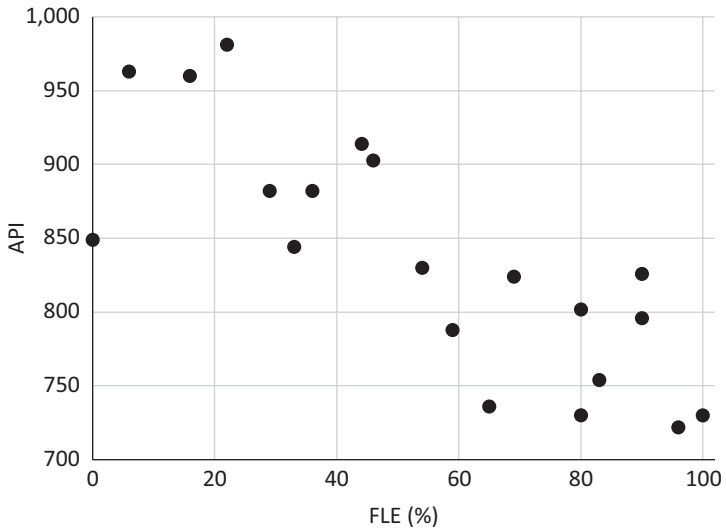

Figure 1.5 API and free lunch eligibility (FLE) for 20 California elementary schools. Academic performance appears to decrease with free lunch eligibility.

In this model, b_1 measures the relationship between free school lunches and schools' academic performance in our sample of 20 schools, b_0 is the intercept (literally, the API corresponding to zero FLE), e_i is an error term that will be much of the focus of this book, and the subscript i denotes school. (Econometricians often use t instead of i as a subscript. We can use any subscript we wish, as long as we're consistent. Often, econometricians use t to refer to time and i to individuals.)

We now use Roman letters like b instead of Greek letters like β to indicate that we are going to estimate this equation from sample data. This is just like using $\bar{X}$ to denote the sample mean estimated from a population with mean μ.

We want to use our data to estimate this simple regression model. Before we can do that, though, we need formulas to compute b_0 and b_1; that is, we have to derive our estimators. That's the subject of the next chapter.

What We Learned in This Chapter

- Econometrics is about making sense of economic data.
- Three steps to conducting econometric analysis.
  1. State the purpose of the analysis.
  2. Formulate the research design and specify the econometric model.
  3. Apply statistical theory.