1

Introduction to Econometrics

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

—Data, Star Trek

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 skies, the weather, migrations of prey, growth of crops, the spread of pestilence. We have evolved to have 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: knowing where to find fish in the sea (and being smart enough to get off the sea when a cold northwest 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 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 need to understand the unseen processes that create these data. For example, we see changes in consumers' income (X) and

demand (Y) from one day to the next. Somewhere out there is a process hidden away in nature, the "true model" that generates Y from X. We're unlikely ever to discover that process completely, but we can use our economic data, theory and statistics to learn something about it. Even if we don't discover every detail of the data generating process, we can find regularities embedded in it. The great renaissance sculptor Michelangelo once said: "I saw the angel in the marble and carved until I set him free." We may never set the angel free, but we can learn a lot about it.

The first step in doing this is to have a theory of how X explains Y. Without a theory of where our data came from we have nowhere to begin. That's why economic theory is the starting point for any econometric analysis. Your best sources of theoretical insights are the intermediate theory courses that are prerequisites for this class. Theory is the key to being able to posit relationships between explanatory variables (e.g., income) and outcomes (demand). It tells us which economic variables should be in the model we are going to estimate. Throughout this course, we will constantly refer to what we learned in our theory courses as the starting point for building econometric models.

Once we have a theoretical model, we need to know its mathematical form to estimate it using our data. Does Y increase linearly with X? Is this relationship quadratic instead of linear? Logarithmic? Sometimes our economic theory can tell us something about the mathematical form of the equation. For example, our micro theory of the firm posits that output does not increase linearly with individual inputs: there are typically decreasing marginal returns. This means that the process generating the data we see on output and inputs is not a linear production function, but a nonlinear one (for example, Cobb-Douglas). Given data on output and inputs, we probably want to estimate a production function with a mathematical form that reflects these nonlinearities. (A Cobb-Douglas production function, we shall see, can be estimated using a log-log form, that is, by taking the log of both the dependent and explanatory variables.)

In short, before we can use our statistical tools to estimate the model, we need both economic theory and mathematics to come up with the model we want to estimate. Econometrics is challenging because it integrates all three of these fields into one.

An Example: Advertising and Sales

Businesses in the United States spent $131 billion on advertising in 2010, according to the advertising intelligence firm Kantar Media.[1] Was it worth it? What impact does ad spending have on sales? More precisely, what is the expected economic return from an additional $1 spent on ads?

A unique research partnership gives you access to a social networking company's sales data. (The company has its reasons to collaborate with you: it's thinking about increasing its ad spending and wants to make sure it will make a significant return on its investment.)

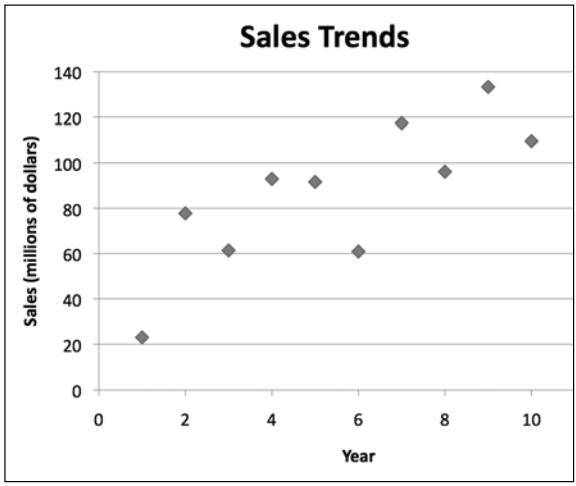
That's the good news. The bad news is that the data will come to you in bits and pieces. You'll have to tell the company exactly what data you'll need in order to do the analysis.

Year	Sales (Y, millions of dollars)
1	23
2	78
3	61
4	93
5	92
6	61
7	117
8	96
9	133
10	110

To kick things off, the company's CEO hands you data on the firms' sales over the last ten years. (These and other data used in this book are available in the online appendices at rebeltext.org.)

Your challenge is to figure out what's driving the firm's sales and whether spending more on advertising boosts sales. Should the company spring for those big Superbowl ads next year? The CEO tells you the company's ad spending has varied from year to year, and she'll have those data to you first thing in the morning.

In the meantime, you can get some mileage out of the stats courses you've taken. You plot sales over time:



There are lots of ups and downs but an overall upward trend. The average is 86 with a variance of 1,027 and a standard deviation of 32.05 (You should use Excel to verify this). You look at recent sales, which are well above the average—as high as 133 million dollars. You set up a t-test and find the odds of having sales that high from a distribution whose mean is 86 and standard deviation 32.05 are very small: less than 10% (120 is 1.46 standard deviations above the mean; again, you should verify this using Excel.)

We all know what's going on here: average sales are increasing over time. The distribution of sales is shifting up to the right. Without this shift

over time, a model for sales might look like this:

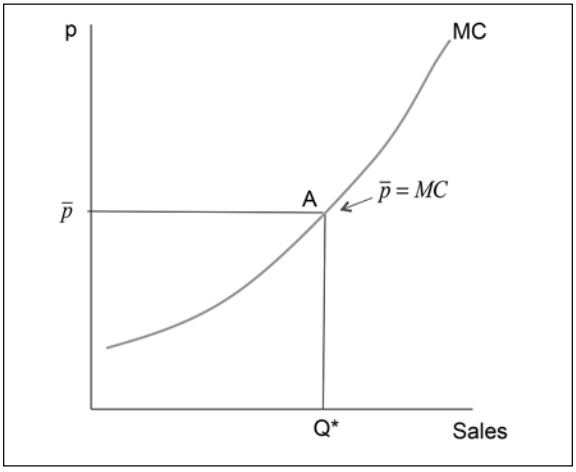
$$Y_i = \mu + \varepsilon_i$$

…where nature takes the "true" population mean of Y (which we call μ, a constant) and then throws in an error, ε, to give us the outcome of Y that we observe. But the mean is not constant; something is shifting it up over time. Some variable, X_i:

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

Notice what we've done here: we've replaced μ with $\mu_i = \beta_0 + \beta_1 X_i$. Mean sales are no longer constant; they depend on X. X could be advertising. It could be that the firm has been spending more on advertising lately, and that's what drove sales. You need a sales model that incorporates advertising.

While you wait for the advertising data, you do some research on the microeconomics of the firm to figure out what your sales model ought to look like. Think of sales as the firm's supply. If the firm is competitive, it will choose to produce where the output price just equals the marginal cost, point A in the next figure. That means the output price and everything that's in the marginal cost determine the firm's output and need to be in your model.



Micro theory tells us that the marginal cost curve embodies the production technology, input prices, and fixed inputs (capital in the short-run). This means both input and output prices should be in your model to explain sales, and so should capital. Changes in any of these variables could explain the changes in sales we see over time. Lurking underneath all this is the firm's technology, which also could have changed. You might be worried that things are starting to get complicated, but by the time this course is finished, you'll have a better idea of how to deal with many of these complications.

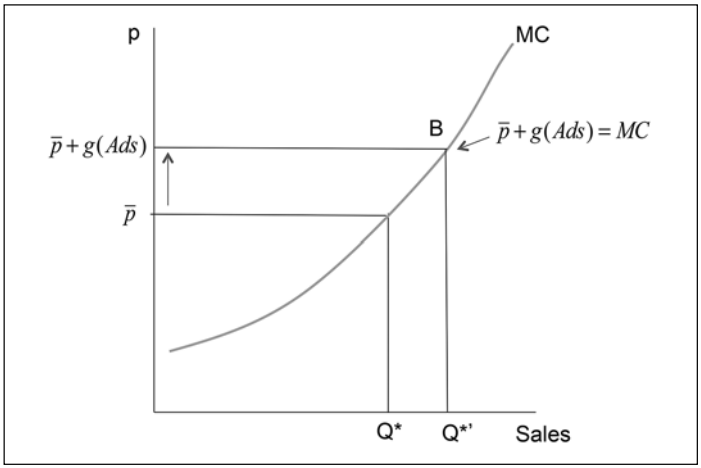
Now stop and think about ads. Those probably weren't mentioned much in your intermediate micro courses, except maybe in the context of shifting out an aggregate demand curve. Competitive firms are price takers in the market.

Where can we find ads in our figure? Are they in marginal costs? Probably not. Ads are a fixed cost in production, not a variable input. The cost of producing that next unit of output will not depend on ads.

Why spend money on ads? Firm-specific ads aim to get consumers to come the firm's way and increase their willingness to pay for its goods; once they've seen the ads, hopefully they'll think what the firm has to offer beats the competition. (Generic ads, like "Every Body Needs Milk," are a little different; they try to shift the whole market demand curve.) The hope is to get a little market power and shift out the demand for the firms' goods.

Where is the demand curve in our figure? It's the willingness-to-pay, or price line. For a competitive firm producing a good exactly like everyone else's, the price line is flat. You get the same price as any other firm for the goods you produce. The goal of your firm's advertising is to increase willingness to pay for your firm's goods and shift the price line up, as shown in the next figure:[2]

In this simple model, advertising increases consumers' willingness to pay for the firm's product. You could think of the price as having two parts: a given market price for an undifferentiated, homogeneous good,



$\bar{p}$, and a price premium consumers are willing to pay for the firm's product. The company hopes its ads will convince people to pay a higher premium. That is:

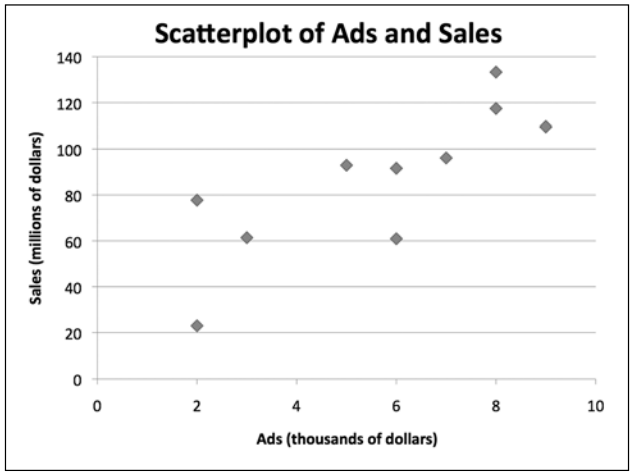
$$p = \bar{p} + g(Ads)$$

Now we have a model, grounded in micro theory, and a hypothesis to test, i.e., that the change in price with respect to ad spending is positive (in the language of calculus, $g'(Ads) > 0$). If ads create a price premium, then one look at our figure is enough to tell us they'll increase sales. We have no idea by how much, though. Our model tells us the firm's sales are a function of output and input prices, capital, and ads:

$$Q_t = F(\bar{p}, w_t, K_t, Ads_t)$$

An increase in sales can therefore result from changes in any of these variables, not only advertising. Ideally, your model will include data on all of these variables in addition to advertising data. We'll talk about models that incorporate multiple explanatory variables in Chapter 3.

You've got a theoretical model just in time for the ads data to hit your inbox. Ads are but one of the right-hand variables you need, but they're a start. You begin by plotting sales against ads:



It looks like there's a positive relationship. You can't be sure, though. It could be one of those other variables, like prices, that explain what we see. Maybe ads increased and prices fell at the same time—due, perhaps, to new technologies that lowered costs in the industry. It could be that falling prices, not rising ads, are behind the increase in sales. It's dangerous to assume that ads explain what we see, when the observed change could be caused by other variables that have been excluded from our model. That's another way of saying that it's important to get the model right.

For now, though, the only data we have are on ads and sales, and we have to start somewhere to learn econometrics! So let's assume, for now, that ads are the only variable explaining sales; that is, let's fit a model of the form:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

where Y is sales, X is ads, ε is an error term that will be much of the focus of this course, and the subscript t denotes year. (Notice we use t here instead of i. 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. In our case, we have time-series data on the same firm but in different years, so let's use t.)

We want to use the data we have to estimate this simple regression model. Before we can do that, though, we need to know how to estimate β_0 and β_1; that is, we have to derive our estimators, or formulas. That's the subject of the next chapter.

Chapter One Notes

1. http://kantarmediana.com/intelligence/press/us-advertising-expenditures-increased-65-percent-2010
2. If your advertising is super effective, consumers might consider your products to be completely different from everyone else's. This gives you market power; it could even make your firm a monopolist. For a monopolist, the price line is not flat; it slopes downward.