# 2

# Simple Regression

*Everything should be made as simple*
*as possible, but not simpler.*
—Albert Einstein

This chapter introduces the simple regression model, that is, a model with only a single right-hand-side variable. We assume that the true model that generated the data we see looks like:

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

where $Y$ is a dependent variable, that is, a variable whose behavior we hope to explain in relation to an explanatory variable, $X$. This model has two parts: a deterministic part, $\beta_0 + \beta_1 X_t$, which describes a line, and a stochastic or random part, which adds the error $\varepsilon_t$ to this line. If you plot your data, with your dependent variable on the $Y$ axis and your explanatory variable on the $X$ axis, $\beta_0 + \beta_1 X_t$ describes a straight line running through your data points. The data points are not exactly on this line, and the error term describes the distance from each point to the line. The error is what makes this a statistical or econometric relationship, instead of a mathematical one.

Before we go on, it's important to emphasize something we said in the last chapter: we do not know what the true model is. This means we do not know the true regression line, $\beta_0 + \beta_1 X_t$, and because we don't

know it, we don't know what the true errors are, either. All we have are the data points. In this book, whenever you see a Greek letter, it will refer to the true value of the thing, which we do not know and therefore must estimate.

Our first job is going to be to find a line that comes closest (we'll discuss what we mean by "closest") to going through the data points. It will be our estimated regression line, and the vertical distance between it and each data point will be the estimated error, or residual. But before we figure out how to estimate this regression line, we need to think a little more about what that true model that gave us our data looks like. For starters, we'd better assume it really *was* a line. We'll drop that assumption soon enough; in Chapter 3 we'll let it be a plane (2 right-hand variables) or a hyperplane (more than 2), and in Chapter 7 we'll let it be all sorts of things.

Next, we need an assumption about how the errors got thrown in when the data were created. For now, we assume that $X$ is not random: there is no error in the $X$ direction, so the error term measures only the vertical distance between the regression line we estimate and the actual data point. (We'll drop that assumption in Chapter 11.)

Imagine an agronomy experiment, in which you plant N identical boxes with the same plant and vary only the amount of fertilizer each box gets. The amount of fertilizer ($X$) applied to each box is given. If Box 3 is supposed to get three pounds of fertilizer, that's what it gets. There is no measurement error. The plant growth you observe ($Y$), on the other hand, is clearly random: if you give two identical boxes exactly the same amount of fertilizer, it is very unlikely that their plant growth will be identical. There will always be an error.

What's in the error? Two things, really. One, which we hope is not important, is the effect of anything we left out of our model that may have varied from box to box. The rest is "white noise," that is, truly random stuff, like random genetic variation from seed to seed. This generally is not a problem for us, especially if there are many boxes!

In a perfectly controlled experiment, left-out variables should not be a problem, assuming these variables are evenly distributed over your boxes. But if one of the boxes got hit by a draft in the night, or temperatures were not controlled for properly, these omitted variables could cause serious problems. Imagine that the boxes that got the most fertilizer just happened to get hit in the night with cold drafts of air, which are not good for growth. It might seem the fertilizer treatment was not effective even if it was. Or the plants that didn't get much fertilizer might have been the ones that caught the draft. In this case, it might seem like the fertilizer was more effective than it really was, because the plants without much fertilizer got stunted by the draft.

Economic behavior is complex, and there are always variables we cannot control for when doing econometric research. The effects of those variables on the outcomes we model naturally fall into the error term. If those missing variables are also correlated with $X$, we could have serious problems, because then $X$ is correlated with the error.

Here's a great example: $Y$ is earnings, $X$ is schooling. Our model (grounded in human capital theory) posits that earnings increase with schooling. (That's a big part of why you are here!) But earnings also depend on innate ability, which is hard to observe. Take two people with the same family background, one driven and entrepreneurial and the other not so much. The first is likely to get more formal education because that's what driven people tend to do. She's also likely to earn more than the second because, well, that's what driven people tend to do. She probably would have earned more even if she did not get the extra education. Our model sees more schooling coinciding with higher earnings and assumes that the schooling, rather than ability, caused the higher earnings. The problem is that ability is not in our model, which means it is in the error.

This is why we need econometrics. In a perfect experiment, the researcher gets to decide how much fertilizer goes into each box, and ideally, controls the experiment to limit the interference of other vari-

ables that might affect growth. In the real world, people decide how much education to complete, and this decision might be determined by some external factor (like ability) that determines both schooling and earnings. So is it the education or ability that explains higher earnings? Mark Twain once said: "I've never let my school interfere with my education." He might have had a point. Lately, labor economists have worked hard to find ways to control for ability while modeling people's earnings. We'll learn about some ways to deal with this problem later, but for now we'll consider situations where $X$ is not correlated with the error.

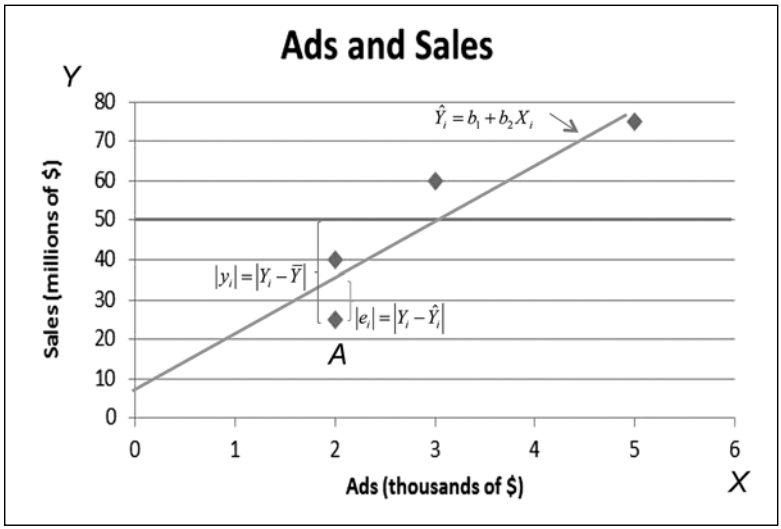
Another requirement of our data is that we have variation in our variables. Since two different points are required to define a line, $X$ must take on at least two different values. If all the people in our sample have the same education level, naturally we cannot estimate the effect of schooling on earnings. There has to be some variation in the $X$ direction. In short, to do what we are about to do we have to assume: (1) that the true relationship between $X$ and $Y$ is linear, and (2) that $X$ is measured without error and takes on at least two different values.

## The Least-Squares Criterion

Consider the scatter plot of ads and sales data in the last chapter. Our goal is to find the line that best fits these data. It will be our best estimate of the true relationship between $X$ and $Y$. We can write the equation for this estimated regression line as follows:

$$Y_t = b_0 + b_1 X_t + e_t$$

Notice that we don't use Greek letters here. We do not know the true parameters, $\beta_0$ and $\beta_1$, so we have replaced them with what we are going to estimate: $b_0$ and $b_1$. We also use $e_t$ to differentiate it from the true error, $\varepsilon_t$, which we do not know. Usually, we use Greek letters to represent things we do not observe, and Roman letters to represent things that we calculate using our data.




Our task is illustrated in the figure above, which for simplicity uses only four data points on ads and sales.

There are many different ways to estimate the relationship between *X* and *Y*. We need a criterion for choosing the best estimator. Overwhelmingly, the one most commonly used in econometrics is what we call the Least-Squares criterion. According to this criterion, the best line is the one that minimizes the sum of squared errors (*SSE*). For each data point, the difference between the observed *Y* and the level of *Y* predicted by the model we will estimate, or the vertical distance between each data point and the regression line, is called the regression error, or residual ($e_t$). The SSE is calculated by squaring each residual then adding them all together. Assuming N observations or data points:

$$SSE = \sum_{t=1}^{N} \left(e_t\right)^2 = \sum_{t=1}^{N} \left(Y_t - b_0 - b_1 X_t\right)^2$$

Our goal, then, is to find the $b_0$ and $b_1$ that minimize this SSE; that is:

$$\min_{b_0, b_1} SSE = \sum_{t=1}^{N} \left( Y_t - b_0 - b_1 X_t \right)^2$$

This is not hard to do using calculus. First, we take the partial derivatives with respect to $b_0$ and $b_1$ and set each one equal to zero (to get the minimum):

$$b_0 : \sum_{t=1}^{N} (Y_t - b_0 - b_1 X_t)(-1) = 0$$

$$b_1 : \sum_{t=1}^{N} (Y_t - b_0 - b_1 X_t)(-X_t) = 0$$

In case you've only had single-variable calculus, taking the partial derivative of a function of more than one variable (here $b_0$ and $b_1$) with respect to one variable is just like taking a simple derivative, while treating the other variable like a constant.

Now at this point you might be wondering whether we've gotten things backwards. In calculus, don't we usually take derivatives of functions with respect to random variables? Didn't we just take derivatives with respect to parameters ($b_0$ and $b_1$) instead?

This is a different use of calculus than you're used to because we're trying to find the parameter values that minimize our SSE function, taking $X_t$ as given. In fact, we'll see soon enough that $b_0$ and $b_1$ *are* random variables. $X_t$, on the other hand, is not random—we've assumed it has no error attached to it whatsoever.

Deriving the two normal equations is the easy part of getting our estimators. The more tedious part is solving the two normal equations

for the two unknowns, $b_0$ and $b_1$. The first equation gives us $b_0$ as a function of $b_1$:

$$\sum_{t=1}^{N}(Y_t - b_0 - b_1 X_t) = 0$$

$$\sum_{t=1}^{N}(Y_t) - \sum_{t=1}^{N}(b_0) - \sum_{t=1}^{N}(b_1 X_t) = 0$$

$$\sum_{t=1}^{N}(Y_t) - Nb_0 - b_1 \sum_{t=1}^{N} X_t = 0$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

where $\overline{Y}$ is the mean of *Y* and $\overline{X}$ is the mean of *X*. (We got the last equation by dividing by N and rearranging.)

We can substitute this in for $b_0$ in the second equation, and after a fair amount of algebraic manipulation we get the formula for $b_1$:

$$\sum_{t=1}^{N}(Y_t - (\overline{Y} - b_1 \overline{X}) - b_1 X_t)(-X_t) = 0$$

$$\sum_{t=1}^{N}\left[(Y_t - \overline{Y}) - b_1 (X_t - \overline{X})\right](X_t) = 0$$

$$\sum_{t=1}^{N}\left[X_t (Y_t - \overline{Y}) - b_1 X_t (X_t - \overline{X})\right] = 0$$

$$b_1 = \frac{\sum_{t=1}^{N} X_t y_t}{\sum_{t=1}^{N} X_t x_t}$$

where small-x and small-y are the deviations of $X_t$ and $Y_t$ from their means: $x_t = X_t - \overline{X}$, and $y_t = Y_t - \overline{Y}$. It turns out (and you should be

able to verify this) that we can write $b_1$ just in terms of deviations from the mean:

$$b_1 = \frac{\sum_{t=1}^{N} x_t y_t}{\sum_{t=1}^{N} x_t^2}$$

Subtracting the mean (a constant) from each $X_t$ subtracts $\bar{X}\sum_{t=1}^{N} x_t$ from both the numerator and denominator, but since the sum of a variable's deviations from its mean is zero, this does not change anything.

These formulas are what we call the ordinary least squares (OLS) estimators of $\beta_0$ and $\beta_1$. Let's use them to estimate our ads and sales model.

## Estimating a Simple Regression Model of Ads and Sales

First, retrieve the ads-sales data. (You should make sure you can derive the OLS estimates presented below in EXCEL; the data are available online, in Appendix 2. You will get some rounding error if you use the data in the book tables to do these calculations.) We need to transform both variables into deviations from their means; see columns D and E of the spreadsheet on the next page.

Then we need $\sum_{t=1}^{N} x_t y_t$ and $\sum_{t=1}^{N} x_t^2$. These are in columns F and G, respectively.

The OLS estimate of $b_1$, which we call $\hat{b}_1$ (we'll use hats "^" to indicate something we've actually estimated), is the sum of column F divided by the sum of column G, or $592/58 = 10.14$. According to our estimates, a $1,000 increase in ads increases sales by $10.14 million. That's a pretty high rate of return! (In a few chapters we'll see that it is too high: remember from last chapter that we suspect advertising is not the only variable that affects sales.) The intercept estimate is $\hat{b}_0 = 86 - (10.14)(6) = 29.61$. In this simple regression model, we can interpret this to mean that, in the absence of any ads spending, expected sales are $29.61 million.

| A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| Year (t) | Sales (Y, millions of dollars) | Ads (X, in thousands of dollars) | y= Y-Ybar | x= X-Xbar | xy | xx |
| 1 | 23 | 2 | -63 | -4 | 228 | 13 |
| 2 | 78 | 2 | -9 | -4 | 31 | 13 |
| 3 | 61 | 3 | -25 | -3 | 65 | 7 |
| 4 | 93 | 5 | 7 | -1 | -4 | 0 |
| 5 | 92 | 6 | 5 | 0 | 2 | 0 |
| 6 | 61 | 6 | -26 | 0 | -10 | 0 |
| 7 | 117 | 8 | 31 | 2 | 75 | 6 |
| 8 | 96 | 7 | 10 | 1 | 14 | 2 |
| 9 | 133 | 8 | 47 | 2 | 113 | 6 |
| 10 | 110 | 9 | 23 | 3 | 79 | 12 |
| Average: | 86 | 6 | | Sums: | 592 | 58 |
| Variance: | 1027 | | | | | |

In short, our estimated regression equation is:

$$\hat{Y}_i = 29.61 + 10.14 X_i$$

Economists often like to present their results as elasticities, which we can also calculate using our estimated regression equation. Recall that the elasticity of $Y$ with respect to $X$ is given by:

$$\eta = \frac{dY / dX}{Y / X}$$

To estimate the elasticity of sales with respect to advertising, we replace the numerator with our estimated regression coefficient, $\hat{b}_1$, which is the slope of our regression line. The elasticity is different depending on where we evaluate it, that is, which values of $Y$ and $X$ we

put in the denominator. Evaluated at the means of *Y* and *X*, the elasticity of sales with respect to ads is:

$$\hat{\eta} = \frac{\hat{b}_1}{\bar{Y} / \bar{X}} = \frac{10.14}{86/6} = 0.66$$

A 1% increase in ads increases sales by 0.66%.

## The R-squared

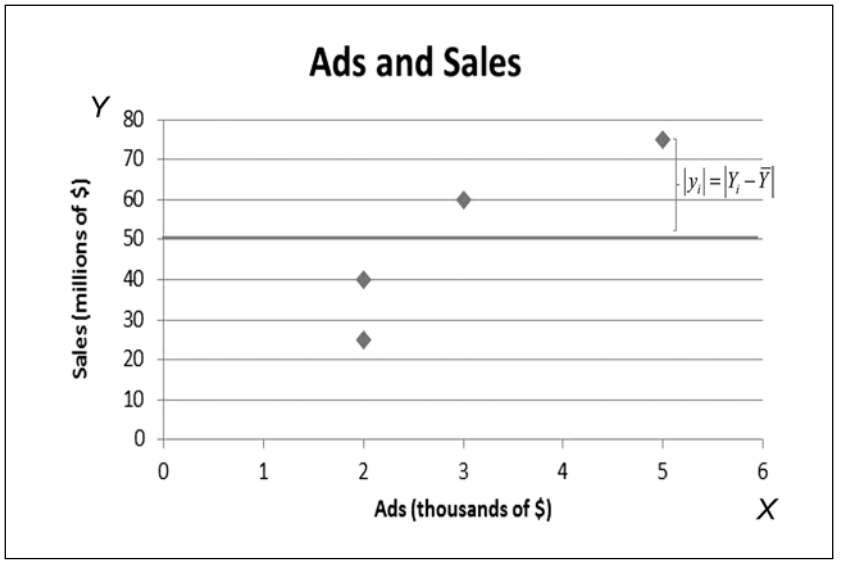
How good is our simple regression model? The goal of our regression is to explain the variation in our dependent variable, *Y*, using the information provided by the independent variable, *X*. One way to measure how well our model fits the data points is to determine how much of the variation in *Y* our model explains. This is what the R-squared statistic tells us.

Let's start by considering the variation in the dependent variable, *Y*. If we take all the deviations of *Y* from its mean, which we have called *y*, square them (so that we have all positive numbers), and add them up, we get the Total Sum of Squares (TSS):

$$TSS = \sum_{i=1}^{N} y_i^2$$

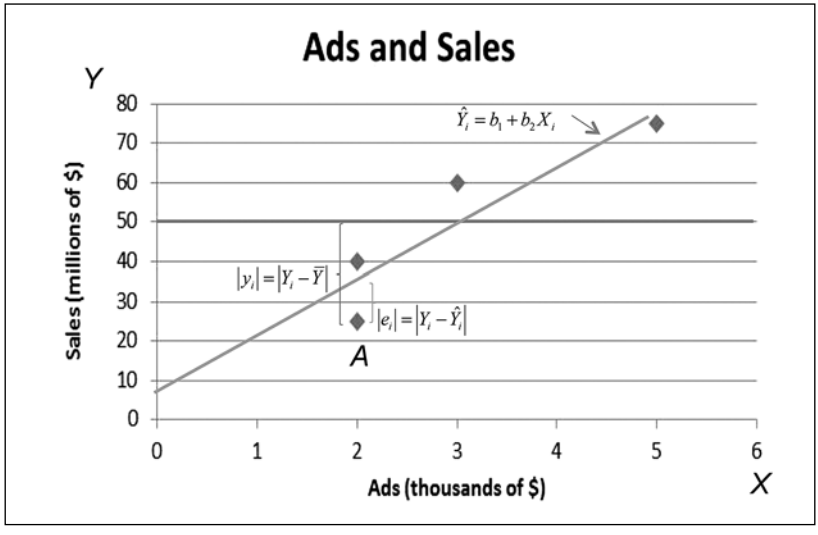
You'll recognize TSS: it's just the numerator in the formula for the variance of *Y* (the denominator is *N-1*). The TSS is unit sensitive: if *Y* is income, its TSS will be 1,000,000 times bigger if we express income in dollars than thousands of dollars! In this sense, it isn't very useful by itself. However, remember that it's the part of the variation in *Y* that is *not* explained by the mean. This will be useful to us later on, because our regression model is "good" if it helps us explain an important share of this variation in *Y* around its mean.

Here's a picture of the variation in *Y* around its mean using our ads and sales example:




As you can see, the values of $Y$ vary quite a bit, and none of them are equal to the mean value. Take the bottom-left point, which corresponds to $(X, Y) = (2, 25)$. We can see that this observed value of $Y$ is way below the mean. Without a regression model, we would only have the mean to work with, and it would not be a very good predictor of $Y$.

Now bring in our regression line:



You can see that our regression does a much better job of predicting *Y* given *X=2* than if we tried to predict *Y* given only the mean value. The estimated regression error, or residual, is small compared to the deviation of *Y* from its mean. In fact, *all* of the data points in this example lie closer to the regression line than to the mean of *Y*. (That won't always be the case, but if you've got a good model, it will tend to be.)

Recall from our derivation of the Least-Squares estimator that if we square all the residuals and add them up, we get the Sum of Squared Errors (SSE). (This is just the numerator in our formula for the regression variance, $s^2$.) This is the variation in *Y* that is *not* explained by our regression, so by dividing the SSE by the TSS (which measures the *total* variation in *Y*), we get the *share* of variation not explained by the regression. Subtract this from 1.0 and we get the share that *is* explained by our regression, or the R-squared:

$$R^2 = 1 - \frac{SSE}{TSS} = 1 - \frac{\sum_{i=1}^{N} \hat{e}_i^2}{\sum_{i=1}^{N} y_i^2}$$

The R-squared statistic is our measure of goodness-of-fit, that is, how good a job our regression does at explaining the variation in *Y* around its mean. If the regression explains little of this variation, our model is not very useful; the SSE will be close in value to the TSS, and $R^2$ will be close to zero. If the regression explains a lot of the variation in *Y* around its mean, the SSE will be small, and $R^2$ will be closer to 1. Because the $R^2$ is a proportion, it will always be the case that $0 \leq R^2 \leq 1$.

So what's a big $R^2$? That depends. Some outcomes are harder to model than others. We usually get a higher $R^2$ from time series regressions than cross-section regressions. There are typically a lot of differences we can't observe across people or whatever it is we are modeling in the cross-section, so we won't be able to explain much of the variation in

the dependent variable given the explanatory variables we can actually observe. In this case, an $R^2$ lower than 0.10 or 0.20 might not be disappointing. If we are modeling the same actor (for example, our firm in the ads and sales example) over time, we are likely to do a better job of explaining variations in the dependent variable. Think of it this way: compare your demand for rice over 52 weeks to 52 people's demand for rice this week. Those 52 people will have all sorts of characteristics you won't be able to observe that might affect their demand for rice. On the other hand, you are you (with the same taste or distaste for rice, etc.) no matter what week it is. Even though we still can't observe all of the characteristics that affect your rice demand, for the most part, they don't change over time, so we don't have to worry about them. It is likely that we'll be able to explain your variations in rice demand better than the variations in rice demand of many different people.

We can calculate the $R^2$ for our simple regression of ads and sales. If we sum the squares of the residuals, we'll get 3237. The previous table reported that the variance of sales is 1027. Remember that the variance is the TSS/(N-1). In our example $N=10$, so the TSS must be 9(1027)=9242. Thus:

$$R^2 = 1 - \frac{SSE}{TSS} = 1 - \frac{3237}{9242} = 0.65$$

Our simple regression succeeds in explaining 65% of the variation of sales around their mean. Not too shabby. But what is explaining the other 35% of the variation? Remember from the introduction that we thought sales would be a function of input and output prices and capital as well as advertising. It may be that, by including variables we've left out in this simple regression model, we can explain more of the variation in sales and even improve our estimate of the relationship between advertising and sales.

## Beyond Simple Regression

Should you present these estimates to your company's board of directors? You probably have doubts about this model, since theory tells us there are other variables that are likely to affect sales. In the next chapter we will consider what happens when we include multiple explanatory variables in our regression model. Then in Chapters 4 and 5 we'll test the statistical significance of our findings and set up confidence intervals around them. In the meantime you might just want to hold off talking to the board!