CHAPTER 5

Regression for Describing and Forecasting

What You'll Learn

- Regression involves finding the line of best fit through some data. It is perhaps the most important tool for describing the relationship between two or more variables.
- Under certain conditions, regression can be useful for forecasting.
- Things can go wrong with regression, especially if we have a small amount of data. Among the most important problems that can arise is overfitting.
- Where did regression come from?

Introduction

In chapter 2, we defined *correlation* and discussed its three uses: description, forecasting, and causal inference. We also talked about a variety of ways to quantify correlations, including the slope of the regression line, the covariance, and the correlation coefficient. Regression lines are the most common and useful of these. In this chapter, we are going to take a deeper dive into regression to make sure we are all thinking clearly about this important technique.

Regression Basics

Let's return to the data on crime and temperature in Chicago that we discussed back in chapter 2. Figure 5.1 reminds you what a scatter plot of that data looks like.

As you can see just by looking at the data, generally speaking, warmer days have more crime. But you sometimes want to be more precise about the relationship. If you worked for the Chicago Police Department and your boss asked you to summarize the relationship between temperature and crime, they probably wouldn't be particularly pleased if you just showed up with this graph. They might want a simple summary of the relationship that's easy to understand and communicate to people making policy decisions. This is where linear regression comes in.

A line of best fit provides just the kind of accessible summary of the relationship between temperature and crime that we are looking for. Such a line, if well chosen, will do two things. First, for any given temperature, the line gives us a reasonable approximation (or prediction) of the amount of crime. And second, as we discussed in chapter 2, the slope of the line tells us something about the sign and magnitude of

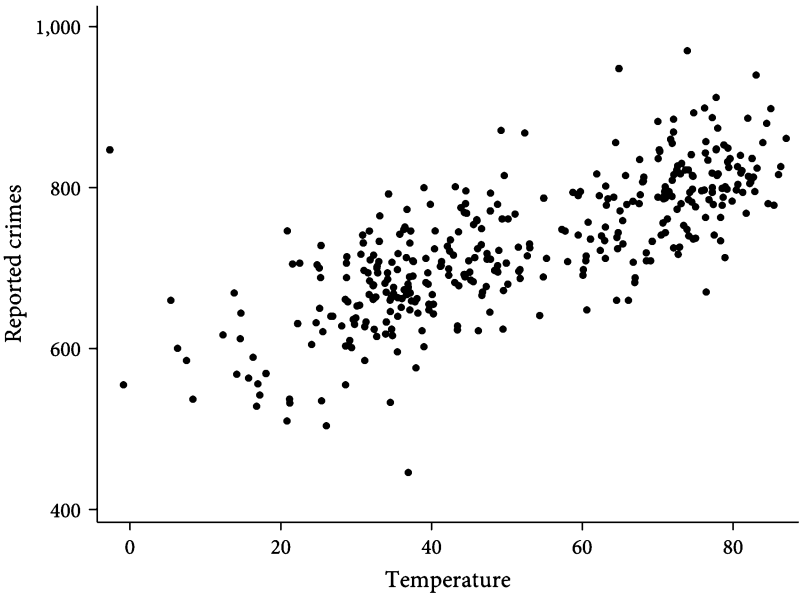

Figure 5.1. Number of reported crimes and the temperature in degrees Fahrenheit in Chicago across days in 2018.

the correlation between the two variables—that is, it tells us approximately how much crime changes as the temperature changes. So let's figure out how we identify the line of best fit, so that we can think clearly about how to interpret and communicate what it has to tell us.

With the exception of completely vertical lines (which wouldn't provide a useful description or forecast anyway), all hypothetical lines that we could draw on the graph of figure 5.1 can be described by what is called a *regression equation* of the following form:

$$\text{Predicted Crime} = \alpha + \beta \cdot \text{Temperature}$$

A regression equation expresses a linear relationship between a *dependent* (or *outcome*) variable on the left-hand side of the equation and an *independent* (or *explanatory*) variable on the right-hand side of the equation. (As we will see later in the chapter, there can be more than one explanatory variable on the right-hand side.) The dependent variable corresponds to the outcome we are trying to describe, predict, or explain. An independent variable corresponds to something we are using to try to describe, predict, or explain the dependent variable.

The regression equation relates the dependent and independent variables linearly through *regression parameters*. The regression parameters define the particular line we are drawing. In our regression equation above, the regression parameters are α and β (the Greek letters *alpha* and *beta*). The regression parameter α is called the *intercept*; it is the predicted number of a crimes on a day when the average temperature is 0 degrees Fahrenheit. The regression parameter β is the *slope*; it is the amount that predicted crime goes up with each degree Fahrenheit. Any possible line on the graph corresponds to one particular combination of α and β. (As we will see later in this chapter, there can be more than two regression parameters if there is more than one independent variable.

And as we will see later in the book, you are free to represent the regression parameters with letters other than α and β when convenient.)

Of course, we don't want to try to describe or predict crime on the basis of temperature using any arbitrary line. The wrong line will yield really bad forecasts. We want to use the line that best fits the data.

In order to find the values of α and β that give us the line that best fits the data, we need to start by defining what the term *best fits* means. We do so quantitatively by choosing a measure of how well any given line does at summarizing or fitting the data. Then we find (or ask our computer to find) the values of α and β that result in the best possible value of that measure. Those values of α and β describe the line of best fit for the data according to the measure we choose.

The measure we choose to evaluate fit is important. As we mentioned briefly in chapter 2, the most commonly used measure (and the one on which we focus) is the *sum of squared errors*. So let's start by being a bit more precise about what this measure means.

For any α and β we choose, our line gives us a prediction of the level of crime on a day with any given temperature. For instance, suppose we chose $\alpha = 650$ and $\beta = 2$. Then, on a day (like January 26, 2018) when the average temperature was 46 degrees Fahrenheit, our prediction of the number of crimes is

$$\text{Predicted Crime} = 650 + 2 \cdot 46 = 742.$$

Of course, the line's prediction won't be exactly right—we sacrifice some accuracy in order to get a parsimonious summary of the data. For instance, in reality, the number of crimes on January 26, 2018, was actually 759. The difference between the true value of the dependent variable and our line's prediction for any given observation is called that observation's *error*:

$$\text{error}_i = \text{Crime}_i - \text{Predicted Crime}_i$$

So, for instance, given our choice of α and β, the error for January 26, 2018, is $759 - 742 = 17$.

Put differently, for any given line we choose (i.e., values of α and β), we can describe any observation i as follows:

$$\text{Crime}_i = \underbrace{\alpha + \beta \cdot \text{Temperature}_i}_{\text{Predicted Crime}_i} + \underbrace{\text{error}_i}_{\text{Crime}_i - \text{Predicted Crime}_i}$$

Figure 5.2 draws a line with $\alpha = 650$ and $\beta = 2$ on top of the data and shows how the errors are measured. (As we will see later, this turns out not to be the line of best fit.) The errors are the vertical lines from a data point to the line. We only drew the errors for a few data points in order to avoid the figure getting too messy. However, to evaluate the fit of a line, we would actually start by calculating the error for every single data point.

The error for any given data point can be positive (if the data point lies above the line) or negative (if the data point lies below the line). But we want a measure of how far the data point is from the line. We don't care whether it is above or below. So, to get such a measure, we next square the error for each data point. The squared error is positive regardless of whether the data point lies above or below the line. It is just a measure of how far the data point is from the line.

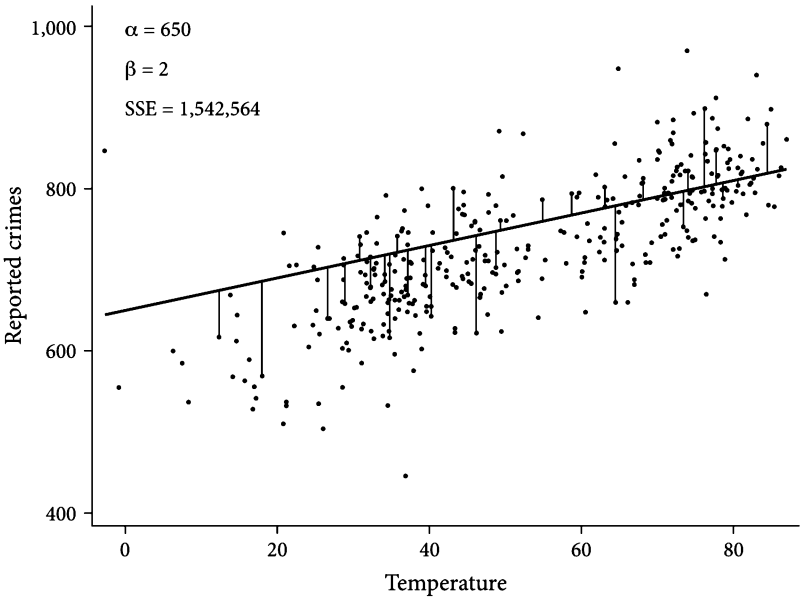

Figure 5.2. Fitting a line through crime and temperature (in degrees Fahrenheit), showing some of the errors.

We add up all those squared errors to get the *sum of squared errors* (SSE). The figure reports the SSE for this particular line in the upper left-hand corner.

We can follow this procedure to get the sum of squared errors for the line associated with any α and β. Different lines have different SSEs. The bigger the sum of squared errors, the further the data is from the line, on average.

The line we are looking for is the one with the smallest sum of squared errors. That is, we find (your computer knows how) the values of the parameters α and β that minimize the sum of squared errors. This process is call *ordinary least squares* (OLS) regression. We label the values of the parameters that minimize the sum of the squared errors as α^{OLS} and β^{OLS}. These values of the parameters are called the *ordinary least squares (OLS) regression coefficients*. The line associated with these parameters is the *OLS regression line*. It is our line of best fit.

There's a lot of lingo to describe finding the α and β that minimize the sum of squared errors. Sometimes we say that we're "regressing crime on temperature." When we're in a long-winded mood, we'll say that we're "running an ordinary least squares regression where crime is the dependent variable and temperature is the independent variable."

Figure 5.3 shows the crime and temperature data with four different lines drawn through it, corresponding to different combinations of α and β. For each line, the figure reports the α, β and the sum of squared errors. A few of the errors are shown visually with vertical black lines. The bottom-right panel shows the OLS regression line—the line that minimizes the sum of squared errors. Visually, we can see that this line is a better approximation of the data than the other three options. In practice, we don't have to use trial and error to find this line. Instead, we'll ask our computer to do the work for us, and it, using linear algebra, will find the values of α and β that minimize the sum of squared errors before you can blink.

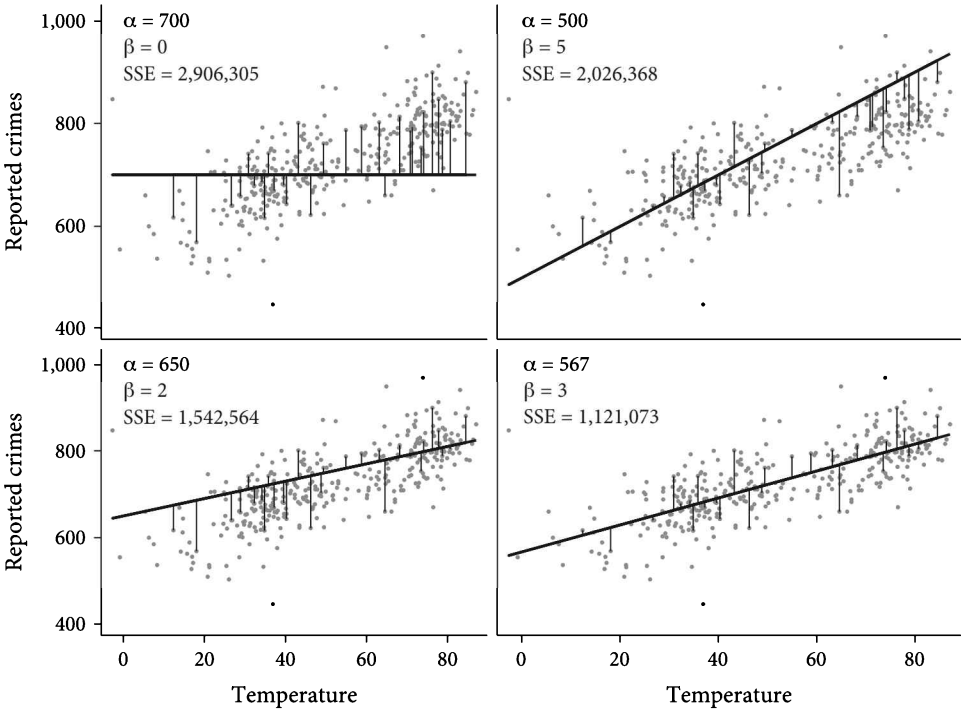

Figure 5.3. Fitting different lines through crime and temperature, showing some of the errors.

How do we interpret the OLS regression line? As we see in the figure, rounding to the nearest integer, the intercept (α^{OLS}) is 567 and the slope (β^{OLS}) is 3. In other words, the OLS regression line is telling us that, in 2018, on days when the average temperature at Midway Airport was 0 degrees, there were about 567 crimes on average, and for every additional degree Fahrenheit, the average number of crimes increased by about 3. So, for example, the predicted amount of crime on a day when the temperature was 46 degress (like January 26, 2018) is

$$\text{Predicted Crime} = 567 + 3 \cdot 46 = 705.$$

We didn't have to choose our regression line by minimizing the sum of squared errors. Depending on our goals, we could have instead minimized the sum of the absolute value of the errors. Or we could have minimized the sum of errors raised to the fourth power. The possibilities are endless.

We like the sum of squared errors for a couple reasons. First, minimizing the sum of squared errors turns out to provide the best linear approximation to another useful function: the *conditional mean function*. The conditional mean function is a function that tells you the mean (average) of some variable conditional on the value of some other variables. Here the particular conditional mean function we are interested in is the one that gives the mean number of crimes conditional on temperature.

Suppose that for, say, each degree Fahrenheit, you calculated the average number of crimes on days with that temperature and plotted them. That gives you a graph of a conditional mean function—for each degree of temperature, it tells you the average number of crimes. In figure 5.4, the light-gray dots are our raw crime and temperature data and the large black dots are the mean number of crimes conditional on being in a

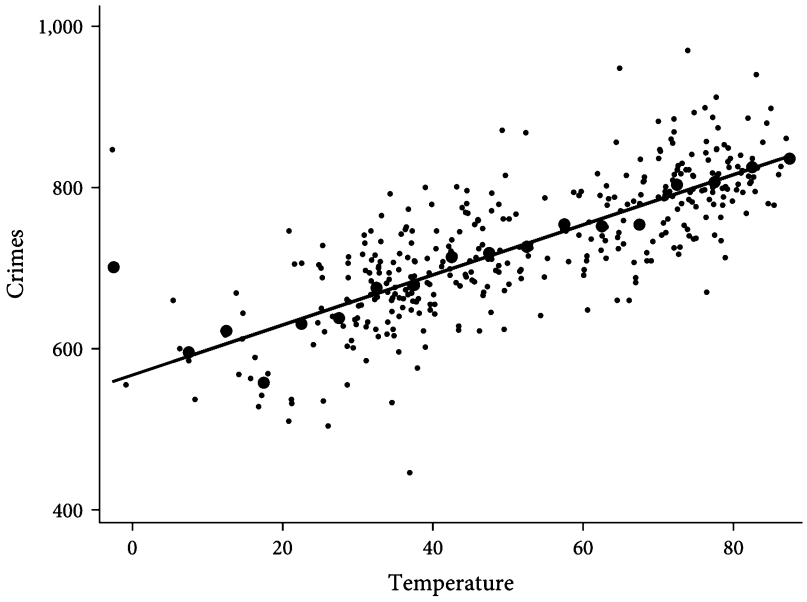

Figure 5.4. The regression line through the data is also the best linear approximation to the conditional means.

5 degree Fahrenheit bin (0–5 degrees, 6–10 degrees, and so on). Conditional means are another reasonable way to predict crime on the basis of temperature. However, the conditional mean function isn't as parsimonious as a line—to summarize the conditional mean function, you need a list of the average level of crime for each temperature bin, whereas a line is summarized by two parameters. But, as you can see, the regression line, in addition to being the line of best fit through the raw data, is also a very good approximation of these conditional means—indeed, it is the best linear approximation of them. So, if you are interested in conditional means, the line that minimizes the sum of squared errors is a good way to summarize them.

Of course, you might not be interested in means. Perhaps, instead, you want to describe or predict the conditional median. In that case, it turns out that you'd want to draw the line that minimizes the sum of the absolute values of the errors. As we said, there are a variety of reasonable choices.

The second reason that people focus on minimizing the sum of squared errors is historical. As indicated above, there is an easy way for your computer to calculate the values of α and β that minimizes the sum of squared errors using linear algebra; as a result, OLS coefficients can be calculated quite quickly. But back when people did this by hand, or even when computers were much slower, this was an important consideration. As computational speeds have improved, however, this consideration has become less relevant.

Linear Regression, Non-Linear Data

What do we do when we want to use linear regression but our data is not well described by a line? To start to think about this, let's return to the data we looked at in our discussion of voter turnout in chapter 2. Remember, there we wanted to describe the relationship between age and voter turnout—perhaps to know whether younger

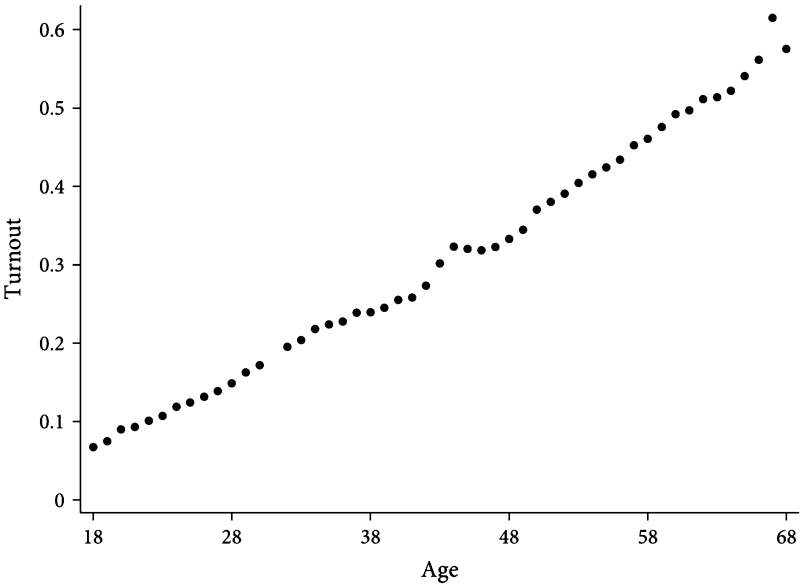

Figure 5.5. Voter turnout rate by age in the 2014 U.S. midterm election.

people are politically underrepresented or perhaps to decide whom to target with a get-out-the-vote drive.

Figure 5.5 shows the voter turnout rate for each year of age between 18 and 68 in the 2014 midterm elections. Notice, in this data, the observation is not an individual; it is an age cohort. As with temperature and crime, the relationship between age and turnout is potentially quite complex. What if we want to summarize the average relationship between age and turnout in a simple way? Or what if we didn't have the data for 31-year-olds (omitted from the figure) and we wanted to come up with our best guess for their turnout level? Or what if we wanted to predict turnout based on age in the 2018 election? Linear regression could be useful for all of these purposes.

Looking at the graph, the relationship between age and turnout appears approximately linear, at least for this range of the data. In other words, we could probably draw a line on this graph that comes pretty close to each of the data points. And if we did draw such a line, this would be fairly useful for both description and forecasting.

Let's try out OLS regression with our voter turnout data. We could again describe any line with the following regression equation:

$$\text{Predicted Turnout} = \alpha + \beta \cdot \text{Age}$$

Our statistical software program tells us that, for this data, $\alpha^{OLS} = -.1381$ and $\beta^{OLS} = .0103$. With these two numbers, we can draw the line that best fits the data, and we can generate predicted turnout for any given age. Figure 5.6 shows how the OLS regression line looks.

It is important to pause and think clearly about the substantive meaning of the regression line.

The number α^{OLS} is the intercept. It tells us that the predicted turnout rate of people age zero is $-.1381$, or about -14 percent. That's a pretty weird prediction. Turnout rates

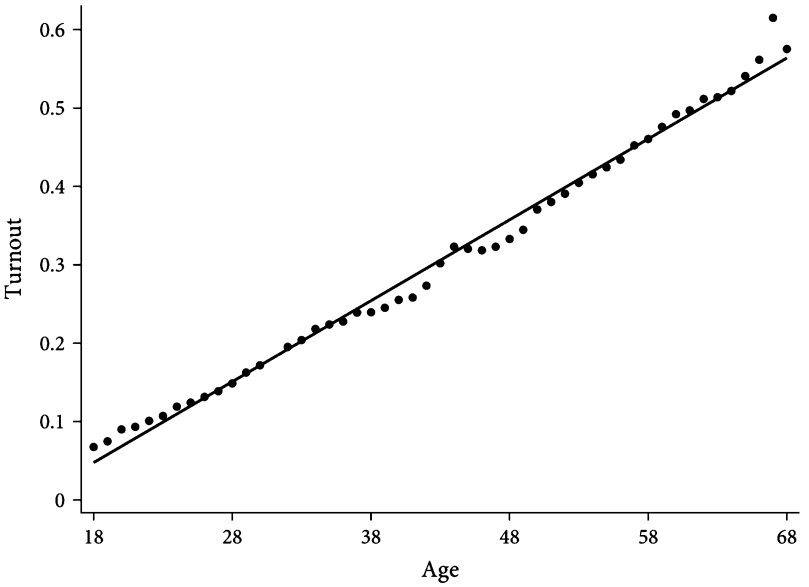

Figure 5.6. OLS regression line through voter turnout rate by age.

can't be negative. And infants can't vote. We know that the turnout rate for zero-year-olds is zero.

Does this mean that the regression is meaningless or wrong? No. It reflects the fact that the regression isn't super useful for describing or forecasting the voting behavior of babies. That isn't surprising. Our regression line was chosen to do a good job of approximating our data. We shouldn't expect it to do a terribly good job approximating the behavior of people with ages well outside the range of our data. And we don't have any data on people younger than 18.

The number β^{OLS} is the slope. It tells us that, on average, within the range of our data, each additional year of age corresponds to an increase in turnout of just over 1 percentage point. In other words, on average, between the ages of 19 and 68, people are about 1 percentage point more likely to vote than people who are just one year younger than themselves. That's interesting. And it accumulates across years, implying that 68-year-olds are approximately 50 percentage points more likely to vote than are 18-year-olds, which is exactly what we see in the data.

The regression line is doing its job pretty well. It gives us a fairly simple and quick summary of the relationship between age and turnout for people between the ages of 18 and 68 in the 2014 election. In this particular election, 18-year-olds voted at an approximate rate of 4.8 percent ($(-.1381 + .0103 \cdot 18) \cdot 100 \approx 4.8$), and then turnout increases by just over 1 percentage point for every additional year in age. Although this summary doesn't get turnout exactly right for each age group, it gets pretty darn close. And, in our view, what is lost in accuracy (compared to, say, just listing turnout rates by age) is more than made up for in parsimony and ease of communication.

We can also use α^{OLS} and β^{OLS} to predict turnout levels for voters whose ages are not in our data. For reasons we've already discussed, we don't want to extrapolate too far. We can't extrapolate back to infants, or even 17-year-olds, since they aren't eligible to vote. We probably also don't want to extrapolate to people too much older than 68.

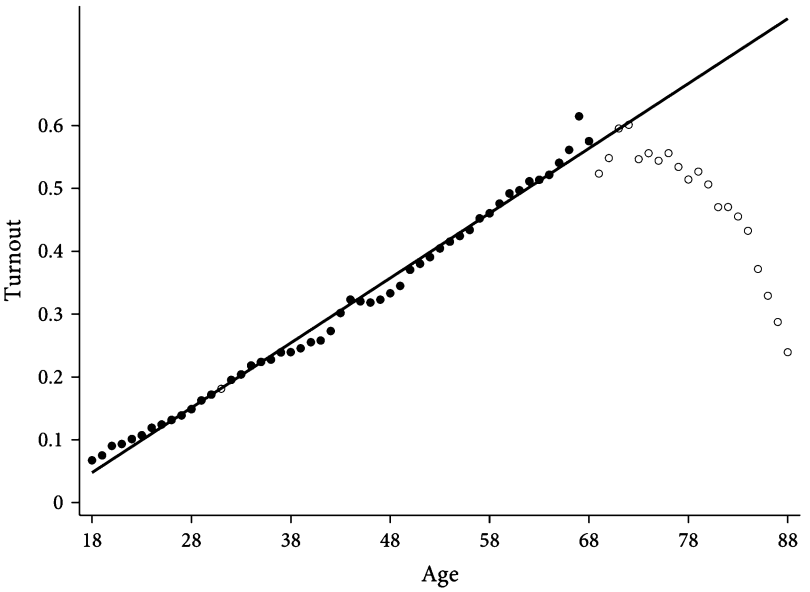

Figure 5.7. Using the regression line to predict voter turnout rate (poorly) for out-of-sample ages.

Our predictions are likely to be pretty good for 69- and 70-year-olds, for whom we predict turnout rates of approximately 57.3 percent $((-.1381 + .0103 \cdot 69) \cdot 100)$ and 58.3 percent $((-.1381 + .0103 \cdot 70) \cdot 100)$, respectively. But the further we get away from the range of our actual data, the more we should worry about the reliability of our predictions.

The spot where we might be most confident in our predictions is for 31-year-olds. For whatever reason, our graph shows no data on that age group. (Here that's because we purposefully omitted it for illustrative purposes. But if you start working with data you'll find that this sort of thing happens all the time. Maybe the county clerk spilled coffee on the voter returns for 31-year-olds.) But we have lots of data on people with ages on both sides of 31. So we can probably generate pretty good predictions for turnout by 31-year-olds. Let's see.

Our regression equation predicts a turnout rate for 31-year-olds of just over 18 percent $(-.1381 + .0103 \cdot 31) \cdot 100 = 18.12$. Since we actually do have the data, we can see how well our prediction pans out by adding the 31-year-olds back in to the graph.

Figure 5.7 plots the same regression line, fit to data on 18- to 68-year-olds, excluding 31-year-olds. But it introduces some previously excluded data points, plotting them as hollow circles. The new data include 31-year-olds, as well as folks ages 69–88.

With 31-year-olds, we hit the mark almost perfectly: we predicted a turnout rate of 18.12 percent, and the true rate was 18.11 percent. With 69- to 72-year-olds, we did okay, though not as well. But our predictions start performing really poorly for the oldest individuals.

That's because the relationship between age and turnout seems to be quite different for the elderly. For younger people, turnout is increasing in age. But once people get past the age of 70 or so, turnout appears to drop with age. As a result, trying to predict the difference in voter turnout between an 80-year-old and an 88-year-old using data on

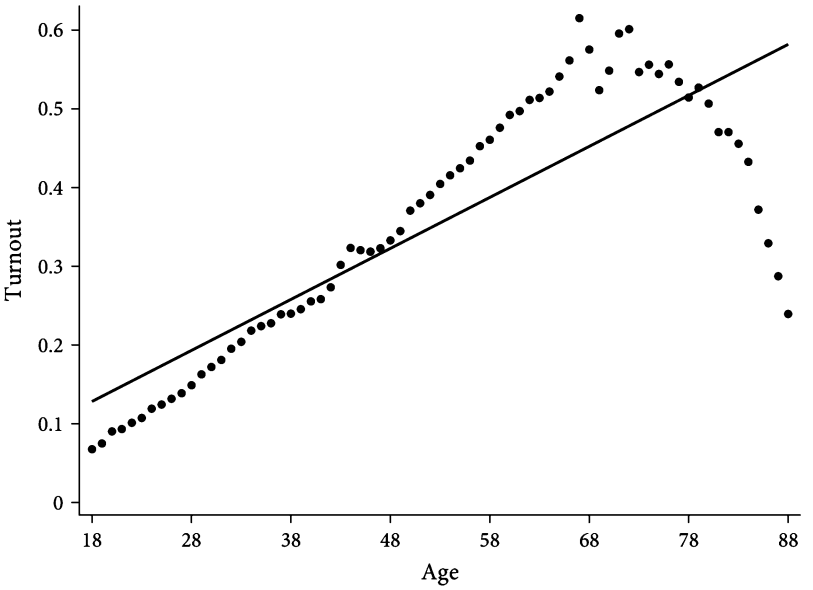

Figure 5.8. A regression of voter turnout rate on age for all ages.

voter turnout of 18- to 68-year-olds doesn't work very well. Just like with our prediction of a −14 percent turnout rate for infants, this illustrates what can go wrong when we try to extrapolate our predictions outside the range of the data that we used to generate the regression line on which those predictions are based.

Suppose we in fact wanted to analyze the relationship between age and turnout for everyone between the ages of 18 and 88. By just looking at the data, we can see that the relationship is not linear. How should we account for this non-linearity?

One approach would be to fit a new linear regression, now using all the data. Even if the data itself doesn't sit on a line, we can still find the line that minimizes the sum of squared errors. As you can see in figure 5.8, there is now a lot more error, since we are fitting a line to data that have a clearly non-linear relationship.

A second approach is to keep fitting regression lines, but use a different line for different parts of the data. For instance, we could find the line that minimizes the sum of squared errors for the data on people between 16 and 68, a second line that minimizes the sum of squared errors for the data on people ages 69–78, and a third line for that data on people ages 79–88. This would not be as parsimonious or easily communicated as running a single regression—instead of two parameters (α and β), we would have six parameters (a separate α and β for each regression line). But, as you can see in figure 5.9, the payoff we get for that lack of parsimony is a tighter fit to the data (i.e., less error).

We hinted at a third way to deal with non-linearity back in chapter 2. There's no reason that our regression equation has to have only one explanatory variable. If we know that there's a non-linear relationship between turnout and age, maybe we want to consider transforming the age variable into age-squared, age-cubed, and so on.

When we took this approach in chapter 2, we kept the regression simple. We just regressed the outcome variable on the explanatory variable squared. But we can do something more general than that. Instead of regressing voter turnout on just age or

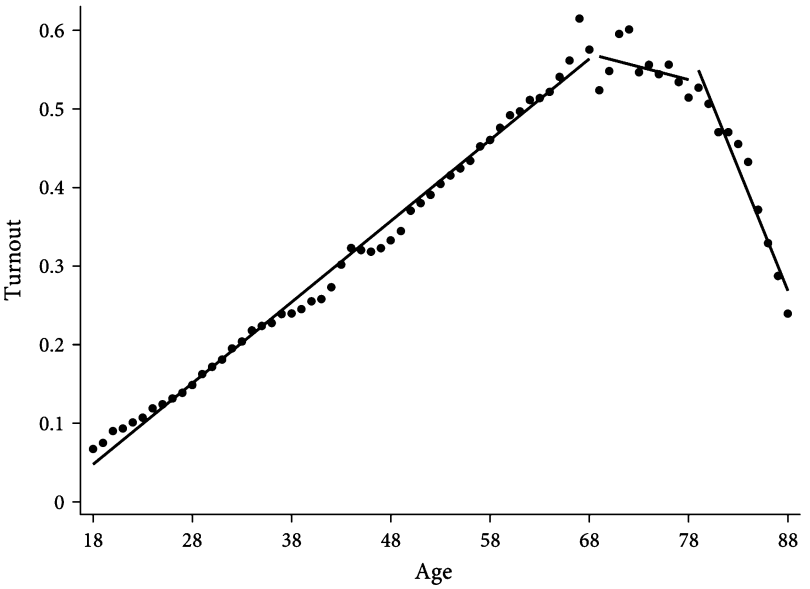

Figure 5.9. Separate regression lines through voter turnout rate and age for ages 16–68, 69–78, and 79–88.

just age-squared, we can regress it on both. This allows us to fit a function that is more flexible than a line to our data. Of course, for each new variable we include, that's a new coefficient that we have to vary when we minimize the sum of squared errors. But our computer can handle that.

In principle, we don't even have to restrict ourselves to different transformations of the age variable. We could also include other factors—average income or average voter registration status—which might further improve our predictions. We'll come back to that possibility in chapter 10. For now, let's stick to transformations of our age variable.

With just one explanatory variable, it is easy to visualize what we are doing when we run a regression. We are just drawing a line through the data in a two-dimensional space—in particular, the line that minimizes the sum of squared errors.

With two explanatory variables, things are a little more abstract, but still manageable. Now we can think about finding a line going through our data in a three-dimensional space. Just picture adding a third axis coming out of the page toward you in our graphs. That axis will have the scale of the second explanatory variable (perhaps age-squared). Now the data forms a cloud in that three-dimensional space. Regression is still just drawing a line that minimizes the sum of squared errors, but now the line passes through that cloud of three-dimensional data points. Describing this line requires three parameters instead of two: the intercept (α), the slope with respect to changes in the first explanatory variable (we can call this β_1), and the slope with respect to changes in the second explanatory variable (we can call this β_2).

Once we go beyond two explanatory variables, it's hard to visualize the regression line, since most of us can't think in four or more dimensions. But you can analogize. You understand what it means to find the line that minimizes the sum of squared errors with one or two explanatory variables. There is no reason we can't do the same with ten. Certainly your computer will have no trouble calculating the sum of squared errors and finding the OLS regression coefficients.

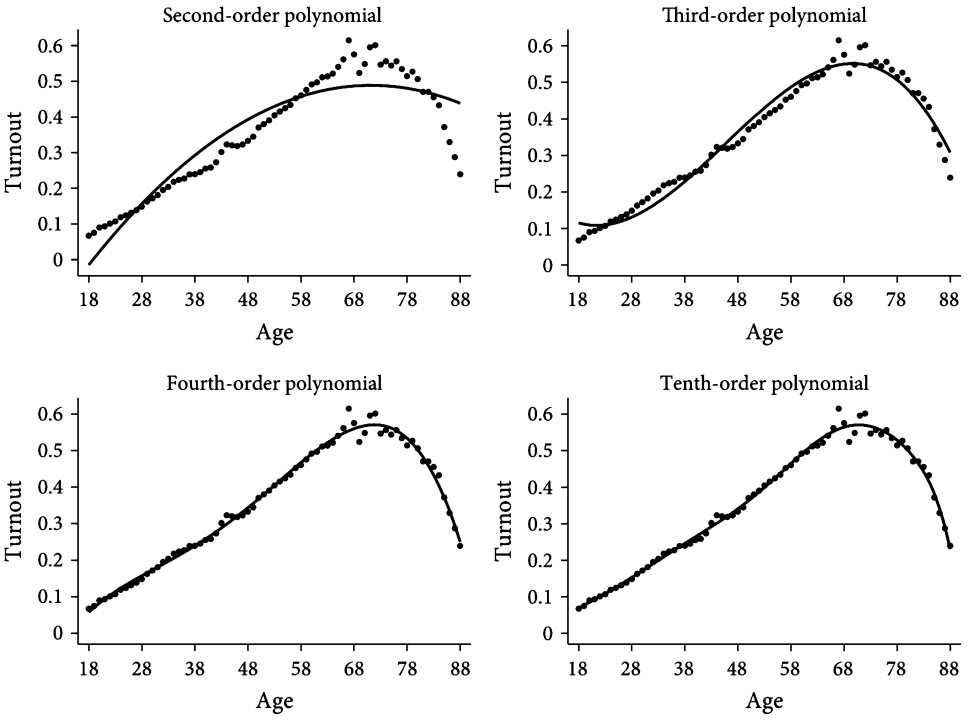

Figure 5.10. Fitting regressions through voter turnout rate with different polynomials of age.

Let's see how this works in practice. We replicated the regression of voter turnout on age but also included age-squared as an explanatory variable. That is, we considered the following equation:

$$\text{Predicted Turnout} = \alpha + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Age}^2$$

Once our computer calculates the associated regression coefficients, we can plug in any value of age and the associated value of age-squared to get a predicted level of turnout. So, for instance, if we wanted to know the predicted turnout of 31-year-olds, we'd plug in 31 for age and $31^2 = 961$ for age-squared.

And we don't have to stop at age and age-squared. Figure 5.10 shows the predicted turnout from different regressions, one with age and age-squared as explanatory variables (this is called a second-order polynomial); another with age, age-squared, and age-cubed as explanatory variables (third-order polynomial); another with a fourth-order polynomial; and another with a tenth-order polynomial!

The overall relationship between age and turnout is pretty complicated. As we've seen, it's approximately linear from 18 to 68, but then it takes a hard turn sometime after that. As a result, if we just include age, we don't do a great job fitting the data. Similarly, we see here that a regression with age and age-squared also doesn't do that well because the relationship in the data is poorly approximated by a quadratic curve. Our predictions get better and better as we include more and more explanatory variables, since we have more and more parameters that we can play around with to fit the data. By the time we get to a fourth-order polynomial, the fit looks quite good.

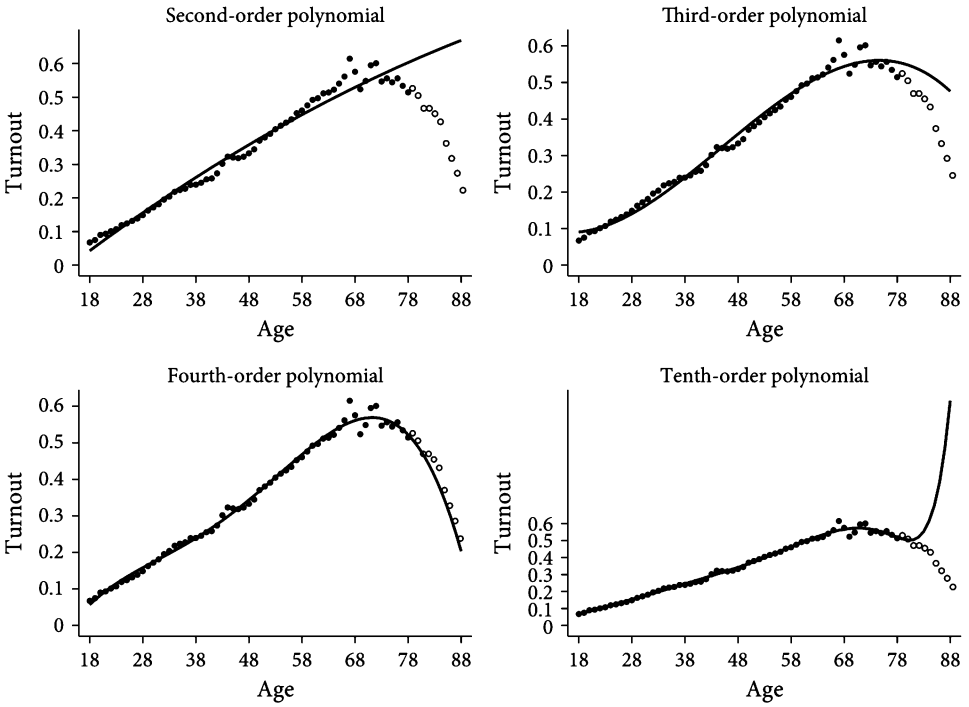

Figure 5.11. Using regressions through voter turnout rate and different polynomials of age to predict voter turnout for out-of-sample ages.

Of course, the tenth-order polynomial does the best job approximating the data—the more explanatory variables included in the regression, the better the fit. But that doesn't necessarily mean that you want to include as many explanatory variables as possible. There are trade-offs.

For one, remember that part of our goal is to describe the data in a simple and parsimonious manner that is easy to understand and communicate. Describing the data with eleven parameters (α plus β_1 through β_{10}) isn't much better in this regard than simply listing the turnout rate for each age group.

Furthermore, we often want to make *out-of-sample predictions*, forecasting voter turnout for age groups not actually observed in our data (like 90-year-olds). Adding more and more terms often results in worse out-of-sample predictions. The reason is that, as the function we use becomes more and more flexible, it can read every little bump and hiccup in the data as meaningful, even when they aren't.

To illustrate this point, we repeated the analyses above, but ran the regressions only using data on people ages 18–78. Then we can see how well we do making out-of-sample predictions of the turnout rates for people with ages above 78. (These predictions are out-of-sample because we purposefully removed voters over the age of 78 from our data.) Figure 5.11 shows the results. The data used for fitting the regression are plotted in black. The data we are attempting to predict are shown as hollow circles. And the gray curve represents the predicted values from the regression. As you can see, the fourth-order polynomial does well at predicting turnout for the oldest voters. But the tenth-order polynomial is a disaster!

The Problem of Overfitting

The example we just saw, where the tenth-order polynomial performed worse than the fourth-order polynomial at out-of-sample prediction, is an instance of a more general phenomenon called *overfitting*. If we test enough explanatory variables, we're bound to find some that correlate with the outcome in our data just by chance. The tenth-order polynomial regression was using meaningless correlations between high-order transformations of the age variable and voter turnout among one set of voters to try to predict turnout among another set of voters. Unsurprisingly, those meaningless correlations did not continue to hold. To better understand overfitting, let's talk about a somewhat more realistic forecasting problem.

Forecasting Presidential Elections

Americans are really interested in predicting the outcomes of upcoming presidential elections. When we tell people that we are political scientists, by far the most common question we get is "Who's going to win the next election?" We tend to disappoint with our answers, since election prediction is not what most political scientists spend their time on.

However, compared to most complex political phenomena, presidential elections are actually rather predictable. Even months before the election, we often have a pretty good idea of who is going to win based on how the economy is doing. And in the final weeks before Election Day, the average of polls usually gets within 1 or 2 percentage points of the final vote share. The journalist Nate Silver established himself as a giant of political data analytics by essentially averaging polls.

Of course, the fact that we can usually predict the vote share within 1 or 2 percentage points doesn't mean we always know who's going to win. Most presidential elections are highly competitive, and the Electoral College allows some candidates to win the election even while losing the popular vote. In close races, like in 2000 or 2016, given the available information on the morning of the election, there was probably no way an honest quantitative analyst could have been more than 90 percent sure that any particular candidate was going to win.

Although we said most political scientists don't spend much time trying to predict election outcomes, some do. The academic journal *PS: Political Science & Politics* typically publishes a symposium before each presidential election with various attempts to predict the outcome using quantitative data and analyses. Often, the goal of these analyses is to see how well researchers can predict the upcoming election results without using polling data. For example, we might see how well we could predict vote share if we just knew the fundamentals, like economic growth and incumbency status.

To make such a prediction, a researcher might run a regression using historical data in which each observation is an election, the outcome variable is the two-party vote share of the incumbent party in that election, and the various explanatory variables are features of that particular election like economic growth in the election year, whether the incumbent is seeking reelection, the number of war casualties over the past four years, and so on. Having obtained the regression coefficients based on data from previous elections, the researcher can then plug in the values for explanatory variables from the current election and obtain a forecast for the upcoming two-party vote share. Because many other analysts are doing the exact same thing, the goal is often to find some new variable to include in your own regression in order to improve its predictive power.

Mimicking this approach, we ran a regression predicting the incumbent's vote share in presidential elections between 1948 and 2012. To be thorough, we included ten different independent variables, all of which have been identified by political scientists as factors that might help us predict election results. Specifically, we included an indicator for whether the incumbent is a Democrat or Republican; an indicator for whether the incumbent is seeking reelection; GDP growth in years 1, 2, 3, and 4 of the most recent presidential term; an indicator for whether the country was involved in a major war at the time; a count of the number of consecutive terms in which the same party has been in power (many people expect that voters are more likely to replace a party that has been in power for a long time); the unemployment rate; and the change in the unemployment rate over the last four years.

We have good reasons to expect that these ten variables should help us predict presidential election results, and at first glance, it looks like they do. The r^2 statistic from the regression is .83, meaning that 83 percent of the variation in incumbent vote share appears to be accounted for by these variables. Furthermore, when we calculate predicted values from this regression, they only miss the actual vote share by an average of 1.7 percentage points.

The apparent success of our regression, however, is misleading. It turns out, if we had simply generated ten random variables (which we have done in computer simulations) and run the same regression using those meaningless numbers as our explanatory variables, we would have averaged an r^2 statistic of around .67 and an average error of 2.4 percentage points. This is almost as good as our predictions using real data, even though our ten randomly generated variables should contain no information about the likely outcome of the election at all.

This is quite surprising. Why is it true? When you generate a bunch of entirely random variables, some of them are going to end up correlated with your outcome just by chance. In a regression, those meaningless variables will appear to predict the outcome. But, of course, they don't really. If you try to use the forecasts generated by the relationship between those meaningless variables and past outcomes to predict future outcomes, you will fail miserably. Their predictive power is just an illusion created by chance.

One way to try to assess and mitigate overfitting is by holding some data out of your regression analysis and conducting out-of-sample tests—as we did with voters over the age of 78 in the previous section. In the context of predicting election outcomes, when generating a prediction for the vote share in 2012, we could leave the 2012 data out of the sample, run a regression using all the other elections, generate a predicted value for 2012 using those regression coefficients and the true values of the explanatory variables for 2012, and see how our predictions fare. In principle, we could do this for each year in our data set—remove one observation, run our regression, generate a predicted value for that observation, check our prediction against the truth, and repeat for each observation.

When we subject our regression with ten explanatory variables to out-of-sample testing, it fares much worse than it first appeared. The average prediction error jumps up from 1.7 to 5.6 percentage points. We doubt that any campaign would hire a statistical consultant who could only promise to predict the election results within 5 or 6 percentage points, on average. Even more embarrassing, a naive prediction based on a simple average of the other elections in the sample gets within 4.6 percentage points, on average. In other words, the overfitted regression that we thought was giving us such accurate predictions is actually worse than a regression that uses no explanatory variables at all.

Table 5.1. Output of regression of average voter turnout on age.

	DV = Voter Turnout
Age	.0103
	(.0001)
Constant	−.1381
	(.0066)
r^2	.991
Root-MSE	.151
Observations	50

Of course, when analysts are careful to avoid overfitting, they can generate useful predictions. A simple regression that uses only GDP growth in year 4 as an explanatory variable produces an out-of-sample prediction error of 3.8 percentage points, beating the model with no explanatory variables. And if we included poll results as Nate Silver does, we would do even better. Nonetheless, it's easy to trick yourself into thinking you're generating good predictions when you're not. Careful analysts only include variables in their regression that they believe are genuinely correlated with the outcome, they avoid having too many variables in their regression relative to the number of observations, and they validate their predictive strategy using out-of-sample testing.

How Regression Is Presented

Sometimes the outputs of a regression are presented graphically, as we have done thus far. But the most common form in which regression results are presented is in a table. For instance, table 5.1 shows how the output of a regression of voter turnout on age might be presented.

You don't quite know what everything in this table means yet (we will come to standard errors, the number in parentheses, in chapter 6), but most everything should be familiar. The number in the Constant row is the intercept, α^{OLS}. The number in the Age row is the slope of the regression line, β^{OLS}. We've also already discussed the idea of r-squared in chapter 2: it is the amount of the variance in voter turnout that can be predicted from age. And Root-MSE is the square root of the mean squared error, which gives you some sense of how far off, on average, our regression predictions are from the real data points.

A Brief Intellectual History of Regression

As far as historians of statistics can tell, regression was invented (or was it discovered?) around the end of the nineteenth century. The first published instance of a linear regression is in the appendix to a brief book entitled *New Methods for the Determination of the Orbits of Comets* by the French mathematician Adrien-Marie Legendre. This was work with important implications for geodesy—the study of the measurement of the earth, which was a high-stakes problem, given the economic and military importance of navigation in the eighteenth century.

Legendre's status as the discoverer of regression was contested by a contemporary, the great German mathematician Carl Friedrich Gauss. In his 1809 *Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*, Gauss staked his claim, writing, "Our principle, which we have made use of since 1795, has lately been published by Legendre." Legendre was not amused, and the two continued to snipe at one another over the matter throughout the early nineteenth century.

Neither Gauss nor Legendre referred to the method of drawing a line of best fit by minimizing the sum of squared errors as a *regression*. That term was coined by the late eighteenth-century scholar, Francis Galton. Galton, a cousin of Charles Darwin's who was also married to Darwin's niece, was a polymath (he dabbled and excelled in a lot of different areas). He also came up with the idea for the modern fingerprinting system and was the first person to quantitatively document the wisdom-of-the-crowds phenomenon.[1] More disturbingly, Galton was a eugenicist—a proponent of selective human breeding. To be clear, we do not support or approve of eugenics, but regression turns out to be useful for non-eugenicists as well.

Galton's interest in eugenics led him to want to study evolution and heredity quantitatively. He started by measuring the easy things like height. In one analysis, he collected data on the heights of parents and their children. After plotting the data, he assessed the average relationship between these two variables using what we now call a regression line.

Galton's analysis was actually a little complicated. He compared the height of children to the average height of their parents after first adjusting the heights so that women's and men's heights were measured on the same scale. We don't want to go through all that. So, to get the idea, imagine an analysis like Galton's that studies just the heights of fathers and sons. The unit of analysis is a father-son pair, and the regression equation looks like this:

$$\text{Predicted Son's Height} = \alpha + \beta \cdot \text{Father's Height}$$

When Galton measured α and β using OLS, what do you think he found? We might have expected $\alpha^{OLS} = 0$ and $\beta^{OLS} = 1$. That would mean, on average, sons tend to be the same height as their fathers—that is, we'd expect the son of a five-foot-tall father to also be five feet tall, the son of a six-foot-tall father to also be six feet tall, and so on. Instead, Galton was surprised to find $\alpha^{OLS} > 0$ and $\beta^{OLS} < 1$. Stop for a moment and think about why that might be.

Figure 5.12 demonstrates Galton's result graphically. The dashed black line shows the 45-degree line—that is, the line with $\alpha = 0$ and $\beta = 1$. The thick gray line shows the best fitting regression line, with $\alpha^{OLS} = 38.2$ and $\beta^{OLS} = 0.448$.

Let's start by interpreting these regression coefficients. The regression line lies above the 45-degree line for relatively short fathers and below for relatively tall fathers. This means that tall fathers tend to have sons that are taller than average but nonetheless shorter than they are. Similarly, short fathers tend to have sons that are shorter than average but nonetheless taller than themselves. Galton called this phenomenon "regression to mediocrity." Today, we typically call this phenomenon *regression to the mean* or *reversion to the mean*, and we'll devote the entirety of chapter 8 to understanding

[1] The idea is that if you ask enough people, even if they are non-experts, perhaps their errors will cancel out and you'll get a good answer. Galton showed that although most individuals are bad at guessing the weight of an ox, if you ask hundreds of people and average their answers, you'll get very close to the correct weight. Unfortunately, this doesn't always work.

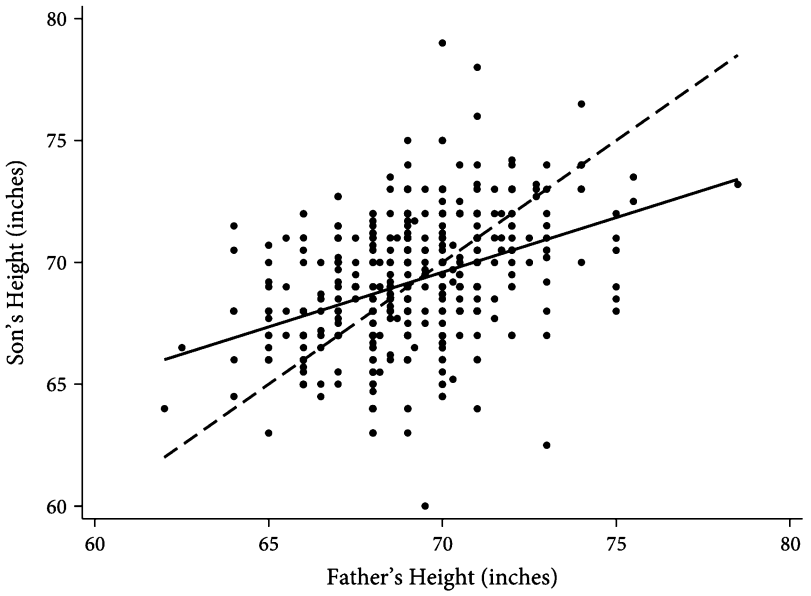

Figure 5.12. A regression line through son's height and father's height.

it. Since then, we've used the word *regression* to refer both to Galton's statistical technique and to the phenomenon that he discovered using it. So it's no coincidence that *OLS regression* and *regression to the mean* use the same word. They have a common intellectual history.

Wrapping Up

Regression is the most important tool we have for studying correlations. The slope of the line of best fit tells us the sign and magnitude of the relationship between two variables—as one goes up, how much does the other tend to go up or down? We can learn a lot from regressions, but we have to be vigilant about keeping our thinking clear. When using a technique that your computer can implement for you, it is easy to become complacent. You can guard against some of the pitfalls by plotting your data, considering the possibility of non-linear relationships, and being careful about overfitting.

A regression tells us the relationship between variables in our data. If we are just trying to describe the data, that is informative all on its own. But often we are trying to do more. For instance, we might be trying to infer the relationship between variables in some larger population from the relationship between those variables in our data, which may only be a small sample of the population. How do we know whether a relationship we found in our data is likely to hold in some larger population? Those concerns are the topic of chapter 6.

Key Terms

- **Dependent variable:** The variable associated with the outcome we are trying to describe, predict, or explain.

- **Independent or Explanatory variable:** A variable we are using to try to describe, predict, or explain the dependent variable.
- **Regression equation:** An equation linearly relating a dependent variable to some independent variables.
- **Regression parameters:** The parameters (intercept and slopes) that relate a dependent variable to some independent variables in a regression equation.
- **Error:** The difference between the value of the outcome variable for an individual data point and the predicted value for that same data point. This is sometimes also referred to as the residual.
- **Sum of squared errors (SSE):** For a given line, calculate the error for each data point by finding its vertical distance from the line. The sum of squared errors for that line is found by squaring each of the individual errors and adding them together.
- **Ordinary least squares (OLS) regression:** The method for finding the line of best fit through data that minimizes the sum of squared errors.
- **Regression line:** The line of best fit through the data that one gets from OLS regression.
- **Intercept:** In the context of a regression, the intercept tells us the predicted value of the outcome when the values of all the explanatory variables are set to 0. This is also referred to as the *constant term*. Sometimes the intercept has a substantive interpretation, but sometimes it doesn't because it doesn't make sense to think about situations where all the explanatory variables are zero (for example, predicted voter turnout for people with an age of zero). In any case, we always include the intercept when we run a regression (except in very unusual circumstances where we know from theory that the intercept should be zero).
- **Conditional mean function:** A function that tells you the mean (average) of some variable conditional on the value of some other variables.
- **Out-of-sample prediction:** Using regression (or another statistical technique) to predict the outcome for observations that were not included in the original data you used to generate your predictions.
- **Overfitting:** Attempting to predict a dependent variable with too many independent variables, so that variables appear to predict the dependent variable in the data but have no actual relationship with it in the world.

Exercises

Download SchoolingEarnings.csv and the associated README.txt, which describes the variables in this data set, at press.princeton.edu/thinking-clearly. This data set gives the average annual earnings for 41- to 50-year-old men in the United States in 1980 at each level of schooling. One observation gives the average earnings (in thousands of dollars) for men with eight years of schooling, another gives the average for those with nine years of schooling, and so on.

5.1 Run a regression with earnings as the dependent variable and schooling as the sole independent variable. Interpret the coefficients.

5.2 Suppose you wanted a parsimonious way to predict earnings using only years of schooling. What would you do?

5.3 Let's dig more deeply into whether the relationship between earnings and schooling is approximately linear.

(a) Start by making a scatter plot. Then plot the predicted values from your regression along with the raw data points, as we did in chapter 2. Does the regression line look like it's fitting the data well?
(b) Now run a fourth-order polynomial regression (i.e., include schooling, schooling^2, schooling^3, and schooling^4. Do those predictions meaningfully differ from the predictions coming from the linear regression?
(c) Now run different regressions for some different ranges of schooling. Do those lines look meaningfully different from the predictions you get from a single regression including all the data?
(d) Does all this make you think the simple linear approach was reasonable or unreasonable?

5.4 Similar to what we did with age and voter turnout, conduct some out-of-sample tests to evaluate your prediction strategy. Using only data for those with twelve years of schooling or less, see how well your different strategies from question 3 perform when predicting earnings for those with more than twelve years of schooling.

Readings and References

For more information on the early history of regression, see

Stephen M. Stigler. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900.* Belknap, Harvard.