CHAPTER 2

Correlation: What Is It and What Is It Good For?

What You'll Learn

- Correlations tell us about the extent to which two features of the world tend to occur together.
- In order to measure correlations, we must have data with variation in both features of the world.
- Correlations *can* be potentially useful for description, forecasting, and causal inference. But we have to think clearly about when they're appropriate for each of these tasks.
- Correlations are about linear relationships, but that's not as limiting as you might think.

Introduction

Correlation doesn't imply causation. That's a good adage. However, in our experience, it's less useful than it might be because, while many people know that correlation doesn't imply causation, hardly anyone knows what correlation and causation are.

In part 1, we are going to spend some time establishing a shared vocabulary. Making sure that we are all using these and a few other key terms to mean the same thing is absolutely critical if we are to think clearly about them in the chapters to come.

This chapter is about correlation: what it is and what it's good for. Correlation is the primary tool through which quantitative analysts describe the world, forecast future events, and answer scientific questions. Careful analysts do not avoid or disregard correlations. But they must think clearly about which kinds of questions correlations can and cannot answer in different situations.

What Is a Correlation?

The *correlation* between two features of the world is the extent to which they tend to occur together. This definition tells us that a correlation is a relationship between two things (which we call *features of the world* or *variables*). If two features of the world tend to occur together, they are *positively correlated.* If the occurrence of one feature of the world is unrelated to the occurrence of another feature of the world, they are *uncorrelated.* And if when one feature of the world occurs the other tends not to occur, they are *negatively correlated.*

Table 2.1. Oil production and type of government.

	Not Major Oil Producer	Major Oil Producer	Total
Democracy	118	9	127
Autocracy	29	11	40
Total	147	20	167

What does it mean for two features of the world to tend to occur together? Let's start with an example of the simplest kind. Suppose we want to assess the correlation between two features of the world, and there are only two possible values for each one (we call these *binary* variables). For instance, whether it is after noon or before noon is a binary variable (by contrast, the time measured in hours, minutes, and seconds is not binary; it can take many more than two values).

Political scientists and economists sometimes talk about the *resource curse* or the *paradox of plenty*. The idea is that countries with an abundance of natural resources are often less economically developed and less democratic than those with fewer natural resources. Natural resources might make a country less likely to invest in other forms of development, or they might make a country more subject to violence and autocracy.

To assess the extent of this resource curse, we might want to know the correlation between natural resources and some feature of the economic or political system. That process starts with collecting some data, which we've done. To measure natural resources we looked at which countries are major oil producers. We classify a country as a major oil producer if it exports more than forty thousand barrels per day per million people. And for the political system we looked at which countries are considered autocracies versus democracies by the Polity IV Project. Table 2.1 indicates how many countries fit into each of the four possible categories: democracy and major oil producer, democracy and not major oil producer, autocracy and major oil producer, and autocracy and not major oil producer.

We can figure out if these two binary variables—being a major oil producer or not and autocracy versus democracy—are correlated by making a comparison. For instance, we could ask whether major oil producers are more likely to be autocracies than countries that aren't major oil producers. Or, similarly, we could ask whether autocracies are more likely to be major oil producers than democracies. If one of these statements is true, the other must be true as well. And these comparisons tell us whether these two features of the world—being a major oil producer and being an autocracy—tend to occur together.

In table 2.1, oil production and autocracy are indeed positively correlated. Fifty-five percent of major oil producers are autocracies ($\frac{11}{20} = .55$) while only about 20 percent of countries that aren't major oil producers are autocracies ($\frac{29}{147} \approx .20$). Equivalently, 27.5 percent of autocracies are major oil producers ($\frac{11}{40} = .275$), while only about 7 percent of democracies are ($\frac{9}{127} \approx .07$). In other words, major oil producers are more likely to be autocracies than are countries that aren't major oil producers, and then, necessarily, autocracies are more likely to be major oil producers than democracies.

As a descriptive matter, we find this positive correlation interesting. It is also potentially useful for prediction. Suppose there were some other countries outside our data

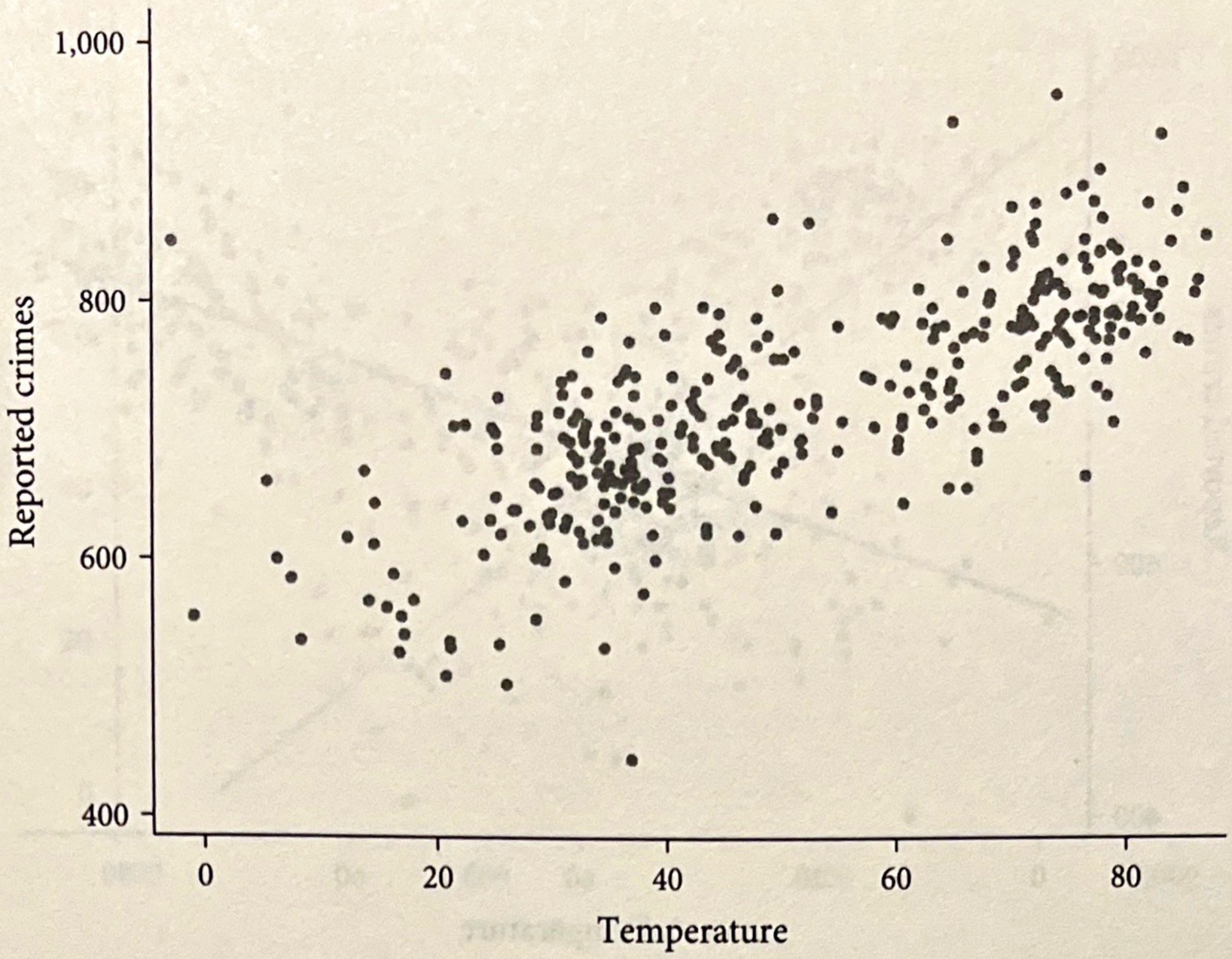

Figure 2.1. Crime and temperature (in degrees Fahrenheit) in Chicago across days in 2018.

whose system of government we were uncertain of. Knowing whether or not they were major oil producers would be helpful in predicting which kind of government they likely have.

Such knowledge could even potentially be useful for causal inference. Perhaps new oil reserves are discovered in a country and the State Department wants to know what effect this is likely to have on the country's political system. This kind of data might be informative about that causal question as well. Though, as we'll discuss in great detail in chapter 9, we must be very careful when giving correlations this sort of causal interpretation.

We can assess correlations even when our data are such that it is hard to make a table of all the possible combinations like we did above. Suppose, for example, that we want to assess the relationship between crime and temperature in Chicago. We could assemble a spreadsheet in which each row corresponds to a day and each column corresponds to a feature of each day. We often call the rows *observations* and the features listed in the columns *variables*. In this case, the observations are different days. One variable could be the average temperature on that day as measured at Midway Airport. Another could be the number of crimes reported in the entire city of Chicago on that day. Another still could indicate whether the *Chicago Tribune* ran a story about crime on its front page on that day. As you can see, variables can take values that are binary (front page story or not), discrete but not binary (number of crimes), or continuous (average temperature).

We collected data like this for Chicago in 2018, and we'd like to assess the correlation between crime and temperature. But how can we assess the correlation between two non-binary variables?

One starting point is to make a simple graph, called a *scatter plot*. Figure 2.1 shows one for our 2018 Chicago data. In it, each point corresponds to an observation in our data—here, that means each point is a day in Chicago in 2018. The horizontal axis of our figure is the average temperature at Midway Airport on that day. The vertical axis

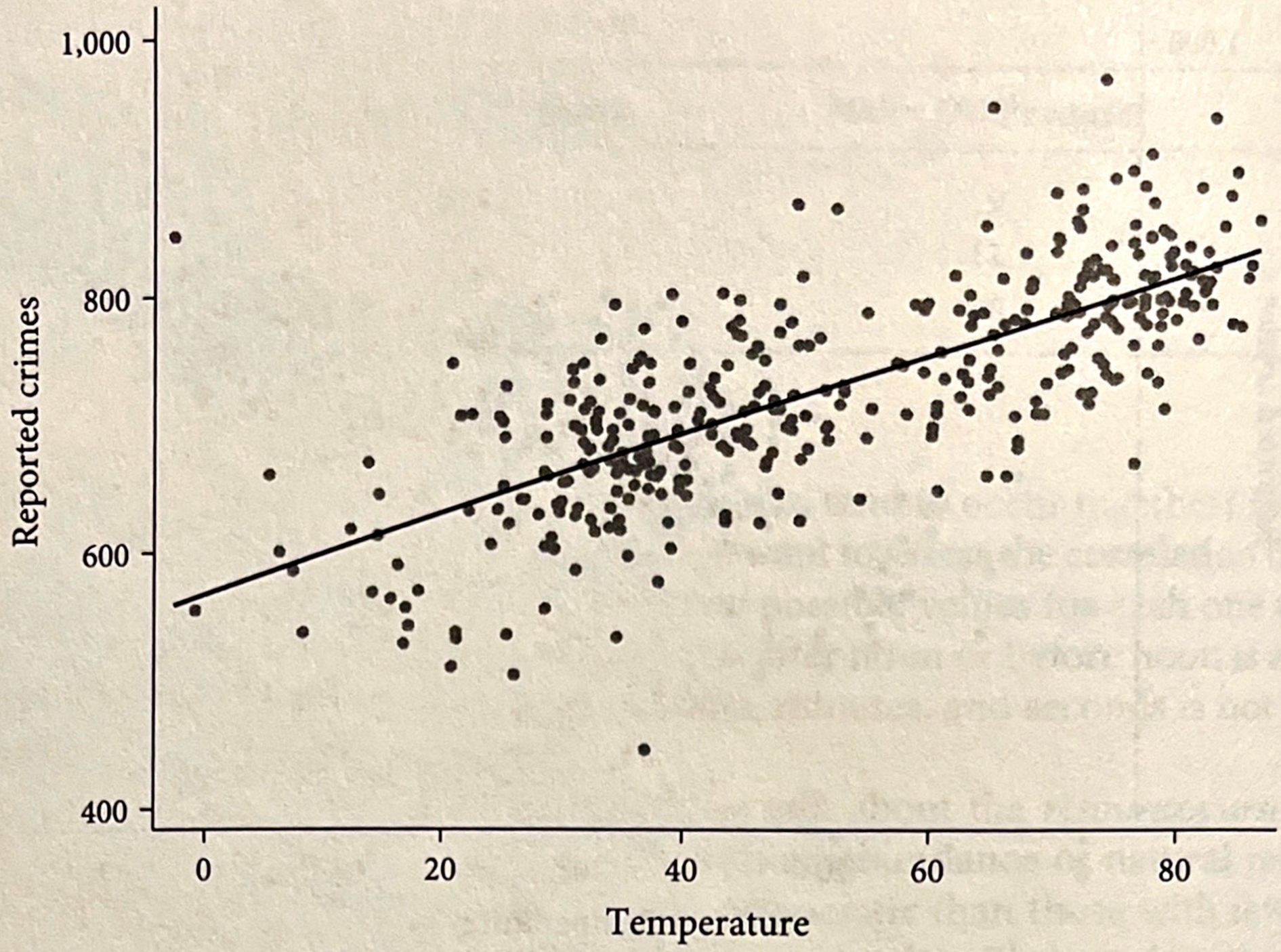

Figure 2.2. A line of best fit summarizing the relationship between crime and temperature (in degrees Fahrenheit) in Chicago across days in 2018.

is the number of crimes reported in the city on that day. So the location of each point shows the average temperature and the amount of crime on a given day.

Just by looking at the figure, you can see that it appears that there is a positive correlation between temperature and crime. Points to the left of the graph on the horizontal axis (colder days) tend to also be pretty low on the vertical axis (lower crime days), and days to right of the graph on the horizontal axis (warmer days) tend to also be pretty high on the vertical axis (higher crime days).

But how can we quantify this visual first impression? There are actually many different statistics that we can use to do so. One such statistic is called the *slope*. Suppose we found *the line of best fit* for the data. By *best fit*, we mean, roughly, the line that minimizes how far the data points are from the line on average. (We will be more precise about this in chapter 5.) The slope of the line of best fit is one way of describing the correlation between these two continuous variables.

Figure 2.2 shows the scatter plot with that line added. The slope of the line tells us something about the relationship between those two variables. If the slope is negative, the correlation is negative. If the slope is zero, temperature and crime are uncorrelated. If the slope is positive, the correlation is positive. And the steepness of the slope tells us about the strength of the correlation between these two variables. Here we see that they are positively correlated—there tends to be more crime on warmer days. In particular, the slope is 3.1, so on average for every additional degree of temperature (in Fahrenheit), there are 3.1 more crimes.

Notice that how you interpret the slope depends on which variable is on the vertical axis and which one is on the horizontal axis. Had we drawn the graph the other way around (as in figure 2.3), we would be describing the relationship between the same two variables. But this time, we would have learned that for every additional

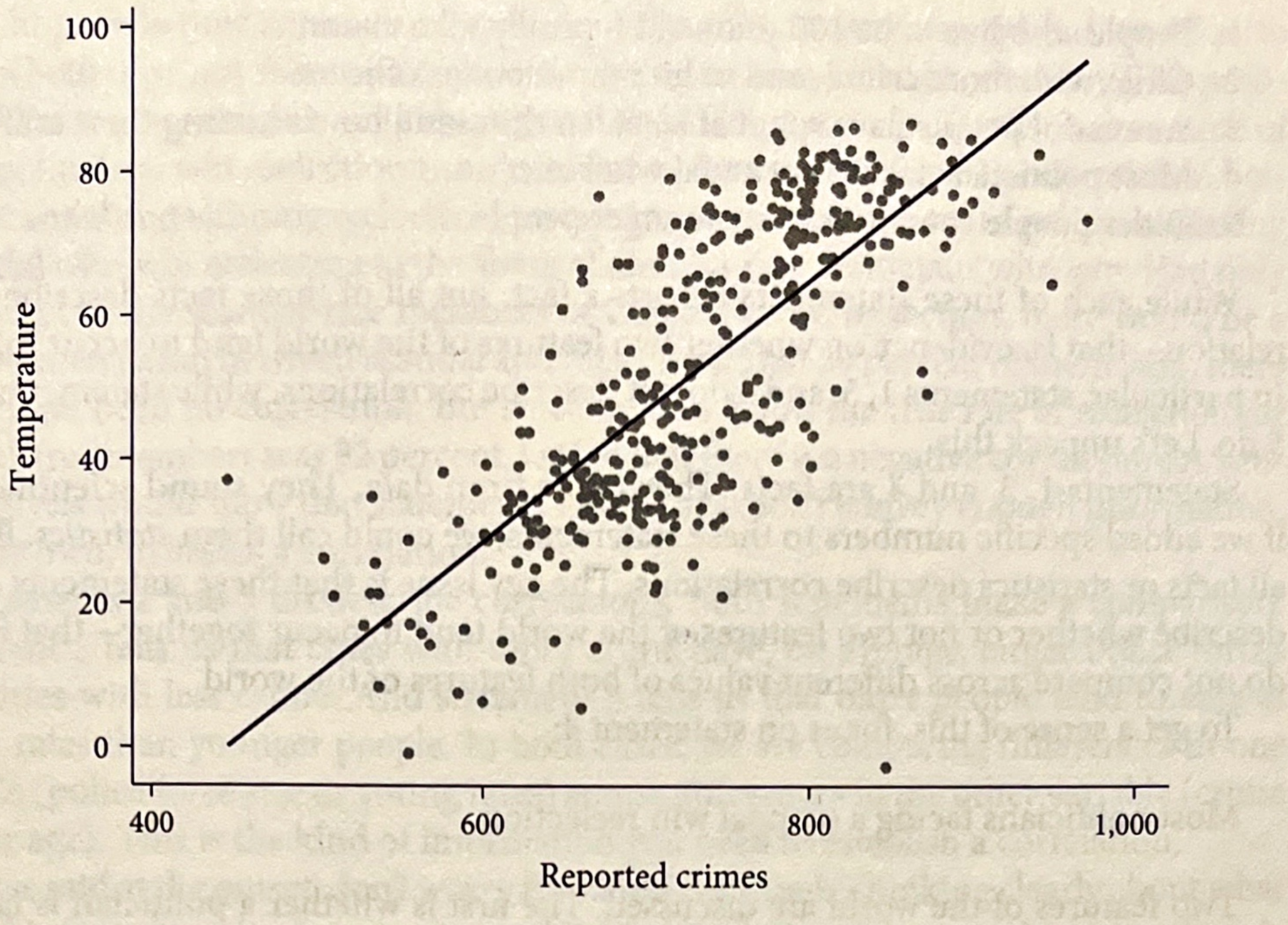

Figure 2.3. A line of best fit summarizing the relationship between temperature and crime in Chicago across days in 2018.

reported crime, on average, the temperature is 0.18 degrees higher. The sign of the slope (positive or negative) is the same regardless of which variable is on the horizontal or vertical axis because changing which variable is on which axis does not change whether they are positively or negatively correlated. But the actual number describing the slope and its substantive interpretation—that is, what it says about the world—has changed.

Fact or Correlation?

In order to establish whether a correlation exists, you must always make a comparison of some kind. For example, to learn about the correlation between temperature and crime, we need to compare hot and cold days and see whether the levels of crime differ, or alternatively, we can compare high- and low-crime days to see if their temperatures differ. This means that to assess the correlation between two variables, we need to have variation in both variables. For example, if we collected data only on days when the average temperature was 0 degrees, we would have no way of assessing the correlation between temperature and crime. And the same is true if we only examined days with five hundred reported crimes.

With this in mind, let's pause to check how clearly you are thinking about what a correlation is and how we learn about one. Don't worry if you aren't all the way there yet. Understanding whether a correlation exists turns out to be tricky. We are going to spend all of chapter 4 on this topic. Nonetheless, it is helpful to do a preliminary check now. So let's give it a try.

Think about the following statements. Which ones describe a correlation, and which ones do not?

1. People who live to be 100 years old typically take vitamins.
2. Cities with more crime tend to hire more police officers.
3. Successful people have spent at least ten thousand hours honing their craft.
4. Most politicians facing a scandal win reelection.
5. Older people vote more than younger people.

While each of these statements reports a fact, not all of those facts describe a correlation—that is, evidence on whether two features of the world tend to occur together. In particular, statements 1, 3, and 4 do not describe correlations, while statements 2 and 5 do. Let's unpack this.

Statements 1, 3, and 4 are facts. They come from data. They sound scientific. And if we added specific numbers to these statements, we could call them *statistics*. But not all facts or statistics describe correlations. The key issue is that these statements do not describe whether or not two features of the world tend to occur together—that is, they do not compare across different values of both features of the world.

To get a sense of this, focus on statement 4:

> Most politicians facing a scandal win reelection.

Two features of the world are discussed. The first is whether a politician is facing a scandal. The second is whether the politician successfully wins reelection. The correlation being hinted at is a positive correlation between facing a scandal and winning reelection. But we don't actually learn from this statement of fact whether those two features of the world tend to occur together—that is, we have not compared the rate of reelection for those facing scandal to the rate of reelection for those not facing scandal.

We can assess this correlation, but not with the data described in statement 4. To assess the correlation, we'd need variation in both variables—facing a scandal and winning reelection. Just for fun, let's examine this correlation in some real data on incumbent members of the U.S. House of Representatives seeking reelection between 2006 and 2012. Scott Basinger from the University of Houston has systematically collected data on congressional scandals. Utilizing his data, let's see how many cases fall into four relevant cases: members facing a scandal who were reelected, members facing a scandal who were not reelected, scandal-free members who were reelected, and scandal-free members who were not reelected.

In table 2.2, we see that statement 4 is indeed a fact: 62 out of 70 (about 89%) members of Congress facing a scandal who sought reelection won. But we also see that most members of Congress not facing a scandal won reelection. In fact, 1,192 out of 1,293 (about 92%) of these scandal-free members won reelection. By comparing the scandal-plagued members to the scandal-free members, we now see that there is actually a slight negative correlation between facing a scandal and winning reelection.

Table 2.2. Most members of Congress facing a scandal are reelected, but scandal and reelection are negatively correlated.

	No Scandal	Scandal	Total
Not Reelected	101	8	109
Reelected	1,192	62	1,254
Total	1,293	70	1,363

We hope it is now clear why statement 4 does not convey enough information to know whether or not there is a correlation between scandal and reelection. The problem is that the statement is only about politicians facing scandal. It tells us that more of those politicians win reelection than lose. But to figure out if there is a correlation between scandal and winning reelection, we need to compare the share of politicians facing a scandal who win reelection to the share of scandal-free politicians who win. Had only 85 percent of the scandal-free members of Congress won reelection, there would be a positive correlation between scandal and reelection. Had 89 percent of them won, there would have been no correlation. But since we now know the true rate of reelection for scandal-free members was 92 percent, we see that there is a negative correlation. A similar analysis would show that statements 1 and 3 also don't convey enough information, on their own, to assess a correlation.

Statements 2 and 5 do describe correlations. Both statements make a comparison. Statement 2 tells us that cities with more crime have, on average, larger police forces than cities with less crime. And statement 5 tells us that older people tend to vote at higher rates than younger people. In both cases, we are comparing differences in one variable (police force size or voting rates) across differences in the other variable (crime rates or age). This is the kind of information you need to establish a correlation.

As we said at the outset, don't worry if you feel confused. Thinking clearly about what kind of information is necessary to establish a correlation, as opposed to just a fact, is tricky. We are going to spend chapter 4 making sure you really get it.

What Is a Correlation Good For?

Now that we have a shared understanding of what a correlation is, let's talk about what a correlation is good for. We've noted that correlations are perhaps the most important tool of quantitative analysts. But why? Broadly speaking, it's because correlations tell us what we should predict about some feature of the world given what we know about other features of the world.

There are at least three uses for this kind of knowledge: (1) description, (2) forecasting, and (3) causal inference. Any time we make use of a correlation, we want to think clearly about which of these three tasks we're attempting and what has to be true about the world for a correlation to be useful for that task in our particular setting.

Description

Describing the relationships between features of the world is the most straightforward use for correlations.

Why might we want to describe the relationship between features of the world? Suppose you were interested in whether younger people are underrepresented at the polls in a particular election, relative to their size in the population. A description of the relationship between age and voting might be helpful. Figure 2.4 shows a scatter plot of data on age and average voter turnout for the 2014 U.S. congressional election. In this figure, an observation is an age cohort. For each year of age, the figure shows the proportion of eligible voters who turned out to vote.

The figure also plots the line that best fits the data. This line has a slope of 0.006. In other words, on average, for every additional year of age, the chances that an individual turned out to vote in 2014 increases by 0.6 percentage points. So younger people do indeed appear to be underrepresented, as they turn out at lower rates than older people.

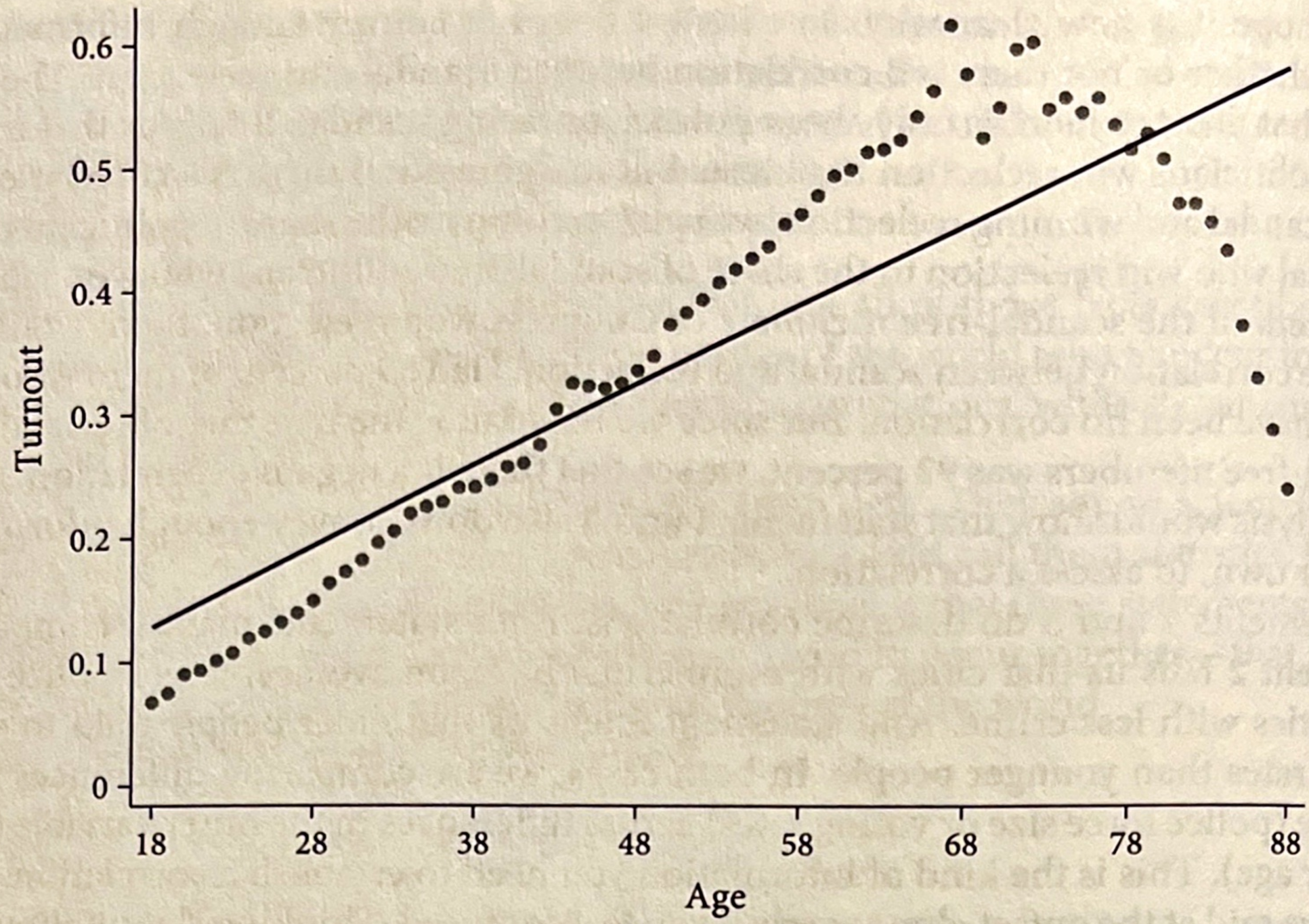

Figure 2.4. Voter turnout and age in the 2014 election.

This kind of descriptive analysis may be interesting in and of itself. It's important to know that younger people were less likely than older people to vote in 2014 and were therefore underrepresented in the electoral process. That relationship may inform how you think about the outcome of that election. Moreover, knowledge of this correlation might motivate you to further investigate the causes and consequences of the phenomenon of younger people turning out at low rates.

Of course, this descriptive relationship need not imply that these younger people will continue to vote at lower rates in future elections. So you can't necessarily use this knowledge to forecast future voter turnout. And it also doesn't mean that these younger people will necessarily become more likely to vote as they age. So you probably can't interpret this relationship causally. This descriptive analysis just tells us that older people were more likely to vote than younger people, on average, in the 2014 election. To push the interpretation further, you'd need to be willing to make stronger assumptions about the world, which we will now explore.

Forecasting

Another motivation for looking at correlations is *forecasting* or *prediction*—two terms that we will use interchangeably. Forecasting involves using information from some sample population to make predictions about a different population.

For instance, you might be using data on voters from past elections to make predictions about voters in future elections. Or you might be using the voters in one state to make predictions about voters in another state. Suppose you're running an electoral campaign, you have limited resources, and you're trying to figure out which of your supporters you should target with a knock on the door reminding them to turn out to vote. If you were already highly confident that an individual was going to vote in the absence

of your intervention, you wouldn't want to waste your volunteers' time by knocking on that door. So accurate forecasting of voter turnout rates could improve the efficiency of your campaign.

Correlations like the one above regarding age and voter turnout could be useful for this kind of forecasting. Since age is strongly correlated with turnout, it might be a useful variable for forecasting who is and is not already likely to vote. For instance, if you were able to predict, on the basis of age, that some group of voters is virtually certain to turn out even without your campaign efforts, you might want to focus your mobilization resources on other voters.

To use the correlation between age and voter turnout for forecasting in this way, you don't need to know why they are correlated. But, unlike if you just want to *describe* the relationship between age and voter turnout in the 2014 election, if you want to *forecast*, you need to be willing to make some additional assumptions about the world.

This raises two important concerns that you must think clearly about in order to use correlation for forecasting responsibly. The first is whether the relationship you found in your sample is indicative of a broader phenomenon or whether it is the result of chance variation in your data. Answering this question requires *statistical inference*, which is the topic of chapter 6. Second, even if you are convinced that you've found a real relationship in your sample, you'll want to think about whether your sample is representative of the population about which you are trying to make predictions. We will explore representativeness in greater detail in our discussion of samples and external validity in chapters 6 and 16.

Let's go back to using age and voter turnout from one election to make predictions about another election. Doing so only makes sense if it is reasonable to assume that the relationship between these two variables isn't changing too quickly. That is, the correlation between age and voter turnout in, for example, the 2014 election would only be useful for figuring out which voters to target in the 2016 election if it seems likely that the relationship between age and turnout in 2016 will be more or less the same as the relationship between age and turnout in 2014. Similarly, if you only had data on age and voter turnout in the 2014 election for twenty-five states, you might use the correlation between age and turnout in those states to inform a strategy in the other twenty-five states. But this would only be sensible if you had reason to believe that the relationship between age and turnout was likely to be similar in the states on which you did and did not have data.

You'd also want to take care in making predictions beyond the range of available data. Our data tell us voter turnout rates for voters ages 18–88. Lines, however, go on forever. So the line of best fit gives us predictions for any age. But we should be careful extrapolating our predictions about voter turnout to, say, 100-year-olds, since we don't have any data for them, so we can't know whether the relationship described by the line is likely to hold for them or not, even for the 2014 election. And we can be sure the line's predictions for turnout by 10-year-olds won't be accurate—they aren't even allowed to vote.

Relatedly, when using some statistic, like the slope of a line of best fit, to do prediction, we need to think about whether the relationship is actually linear. If not, a linear summary of the relationship might be misleading. We'll discuss this in greater detail below.

It is worth noting that, in practical applications, it would be unusual to try to do forecasting simply using the correlation between two variables. One might, instead, try to predict voter turnout using its relationship with a host of variables like gender, race,

income, education, and previous voter turnout. We'll discuss such multivariable and conditional correlations in chapter 5.

Using data for forecasting and prediction is a rapidly growing area for analysts in policy, business, policing, sports, government, intelligence, and many other fields. For instance, suppose you're running your city's public health department. Every time you send a health inspector to a restaurant, it costs time and money. But restaurant violations of the health code do harm to your city's residents. Therefore, you would very much like to send inspectors to those restaurants that are most likely to be in violation of the health codes, so as not to waste time and money on inspections that don't end up improving public safety. The more accurately you can forecast which restaurants are in violation, the more effectively you can deploy your inspectors. You could imagine using data on restaurants that did and did not violate health codes in the past to try to predict such violations on the basis of their correlation with other observable features of a restaurant. Plausibly useful restaurant features might include Yelp reviews, information about hospital visits for food poisoning, location, prices, and so on. Then, with these correlations in hand, you could use future Yelp reviews and other information to predict which restaurants are likely in violation of the health codes and target those restaurants for inspection.

This example points to another tricky issue. The very act of using correlations for prediction can sometimes make correlations that held in the past cease to hold in the future. For instance, suppose the health department observes a strong correlation between restaurants that are open twenty-four hours a day and health code violations. On the basis of that correlation, they might start sending health inspectors disproportionately to twenty-four-hour restaurants. A savvy restaurant owner who becomes aware of the new policy might adapt to fool the health department, say closing from 2:00 to 3:00 a.m. every night. This small change in operating hours would presumably do nothing to clean up the restaurant. But the manager would have gamed the system, rendering predictions based on past data inaccurate for the future. We'll discuss this general problem of adaptation in greater detail in chapter 16.

Forecasting would also be useful to a policy maker who would like to know the expected length of an economic downturn for budgetary purposes, a banker who wants to know the credit worthiness of potential borrowers, or an insurance company that wants to know how many car accidents a potential client is likely to get in this year. The managers of our beloved Chicago Bears would love to predict which college football players could be drafted to increase the team's chances of winning a Super Bowl. But given their past track record, we don't hold out much hope. Data can't work miracles.

It is also worth thinking about the potential ethical implications of using predictions to guide behavior. For instance, research finds that consumer complaints about cleanliness in online restaurant reviews are positively correlated with health code violations. This is potentially useful predictive information—governments could use data collected from review sites to figure out where to send restaurant inspectors. In response to such findings, an article in *The Atlantic* declared, "Yelp might clean up the restaurant industry." But a study by Kristen Altenburger and Daniel Ho shows that online reviewers are biased against Asian restaurants—comparing restaurants that received the same score from food-safety inspectors, they find that reviewers were more likely to complain about cleanliness in the Asian restaurants. This means that if governments make use of the helpful predictive correlation between online reviews and health code violations, it will inadvertently discriminate against Asian restaurants by disproportionately targeting them for inspection. Do you want your government to make use of such

information? Or are there ethical or social costs of targeting restaurants for inspection in an ethnically biased way that outweigh the benefits of more accurate predictions? We will return to some of these ethical issues at the end of the book.

Causal Inference

Another reason we might be interested in correlations is to learn about causal relationships. Many of the most interesting questions that quantitative analysts face are inherently causal. That is, they are about how changing some feature of the world would cause a change in some other feature of the world. Would lowering the cost of college improve income inequality? Would implementing a universal basic income reduce homelessness? Would a new marketing strategy boost profits? These are all causal questions. As we'll see throughout the book, using correlations to make inferences about causal relationships is common. But it is also fraught with opportunities for unclear thinking. (Understanding causality will be the subject of the next chapter.)

Using correlation for causal inference has all the potential issues we just discussed when thinking about using correlation for prediction and there are new issues. The key one is that correlation need not imply causation. That is, a correlation between two features of the world doesn't mean one of them causes the other.

Suppose you want to know the effect of high school math training on subsequent success in college. This is an important question if you're a high school student, a parent or counselor of a high school student, or a policy maker setting educational standards. Will high school students be more likely to attend and complete college if they take advanced math in high school?

As it turns out, the correlation between taking advanced math and completing college is positive and quite strong—for instance, people who take calculus in high school are much more likely to graduate from college than people who do not. And the correlation is even stronger for algebra 2, trigonometry, and pre-calculus. But that doesn't mean that taking calculus causes students to complete college.

Of course, one possible source of this correlation is that calculus prepares students for college and causes them to become more likely to graduate. But that isn't the only possible source of this correlation. For instance, maybe, on average, kids who take calculus are more academically motivated than kids who don't. And maybe motivated kids are more likely to complete college regardless of whether or not they take calculus in high school. If that is the case, we would see a positive correlation between taking calculus and completing college even if calculus itself has no effect on college completion. Rather, whether a student took calculus would simply be an indirect measure of motivation, which is correlated with completing college.

What's at stake here? Well, if the causal story is right, then requiring a student to take calculus who otherwise wouldn't will help that student complete college by offering better preparation. But if the motivation story is right, then requiring that student to take calculus will not help with college completion. In that story, calculus is just an indicator of motivation. Requiring a student to take calculus does not magically make that student more motivated. It could even turn out that requiring that student to take calculus might impose real costs—in terms of self-esteem, motivation, or time spent on other activities—without any offsetting benefits.

The exact mistake we just described was made in a peer-reviewed scientific article. The researchers compared the college performance of people who did and did not take a variety of intensive high school math courses. On the basis of a positive correlation, they

suggested that high school counselors "use the results of this study to inform students and their parents and guardians of the important role that high school math courses play with regard to subsequent bachelor's degree completion." That is, they mistook correlation for causation. On the basis of these correlations, they recommended that students who were not otherwise planning to do so should enroll in intensive math courses to increase their chances of graduating from college.

We'll return to the problem of mistaking correlation for causation in part 3. For now, you should note that, although purported experts do it all the time, in general, it is wrong to infer causality from correlations.

Measuring Correlations

There are several common statistics that can be used to describe and measure the correlation between variables. Here we discuss three of them: the *covariance*, the *correlation coefficient*, and the *slope of the regression line*. But before going through these three different ways of measuring correlations, we need to talk about means, variances, and standard deviations—statistics that help us summarize and understand variables.

Mean, Variance, and Standard Deviation

Let's focus on our Chicago crime and temperature data. Recall that in this data set, each observation is a day in 2018. And for each day we observe two variables, the number of reported crimes and the average temperature as measured in degrees Fahrenheit at Midway Airport. We aren't going to reproduce the entire data set here, since it has 365 rows (one for each day of 2018). Table 2.3 shows what the data looks like for the month of January. For the remainder of this discussion, we will treat the days of January 2018 as our population of interest.

For any observation i, call the value of the crime variable $crime_i$ and the value of the temperature variable $temperature_i$. In our data table, i can take any value from 1 through 31, corresponding to the thirty-one days of January 2018. So, for instance, the temperature on January 13 was $temperature_{13} = 12.3$, and the number of crimes reported on January 24 was $crime_{24} = 610$.

A variable has a *distribution*—a description of the frequency with which it takes different values. We often want to be able to summarize a variable's distribution with a few key statistics. Here we talk about three of them.

It will help to have a little bit of notation. The symbol $\sum$ (the upper-case Greek letter *sigma*) denotes summation. For example, $\sum_{i=1}^{31} \text{crime}_i$ is the sum of all the values of the crime variable from day 1 through day 31. To find it, you take the values of crime for day 1, day 2, day 3, and so on through 31 and sum (add) them together. That is, you add up $\text{crime}_1 = 847$ and $\text{crime}_2 = 555$ and $\text{crime}_3 = 568$ and so on through $\text{crime}_{31} = 708$. You find these specific values for the crime variable on each day by referring back to the data in table 2.3.

Now we can calculate the *mean* of each variable's distribution. (Sometimes this is just called the *mean of the variable*, leaving reference to the distribution implicit). The mean is denoted by μ (the Greek letter *mu*). The mean is just the average. We find it by summing the values of the observations (which we now have convenient notation for) and dividing by the number of observations. For January 2018, the means of our two variables are

Table 2.3. Average temperature at Chicago Midway Airport and number of crimes reported in Chicago for each day of January 2018.

Day	Temperature (°F)	Crimes
1	−2.7	847
2	−0.9	555
3	14.2	568
4	6.3	600
5	5.4	660
6	7.5	585
7	25.4	535
8	33.9	618
9	30.1	653
10	44.9	709
11	51.7	698
12	21.6	705
13	12.3	617
14	15.7	563
15	16.8	528
16	14.6	612
17	14.7	644
18	25.6	621
19	34.8	707
20	40.4	724
21	42.9	716
22	48.9	722
23	32.3	716
24	29.2	610
25	35.5	640
26	46.0	759
27	45.6	754
28	35.0	668
29	25.2	650
30	24.7	632
31	37.6	708
Mean	26.3	655.6
Variance	220.3	5183.0
Standard deviation	14.8	72.0

$$\mu_{\text{crime}} = \frac{\sum_{i=1}^{31} \text{crime}_i}{31} = \frac{847 + 555 + \cdots + 708}{31} = 655.6$$

and

$$\mu_{\text{temperature}} = \frac{\sum_{i=1}^{31} \text{temperature}_i}{31} = \frac{-2.7 + -0.9 + \cdots + 37.6}{31} = 26.3.$$

A second statistic of interest is the *variance*, which we denote by σ^2 (the lower-case Greek letter *sigma*, squared). We'll see why it is squared in a moment. The variance is a way of measuring how far from the mean the individual values of the variable tend to be. You might even say that the variance measures how variable the variable is. (You can also think of it, roughly, as a measure of how spread out the variable's distribution is.)

Here's how we calculate the variance. Suppose we have some variable X (like crime or temperature). For each observation, calculate the *deviation* of that observation's value of X from the mean of X. So, for observation i, the deviation is the value of X for observation i (X_i) minus the mean value of X across all observations (μ_X)—that is, $X_i - \mu_X$. On January 13, 2018, the temperature was 12.3 degrees Fahrenheit. The mean temperature in January 2018 was 26.3 degrees Fahrenheit. So January 13's deviation from the January mean was $12.3 - 26.3 = -14$. That is, January 13, 2018, was fourteen degrees colder than the average day in January 2018. By contrast, the deviation of January 23, 2018, was $32.3 - 26.3 = 6$. On January 23, it was six degrees warmer than on the average day in January 2018.

Note that these deviations can be positive or negative since observations can be larger or smaller than the mean. But for the purpose of measuring how variable the observations are, it doesn't matter whether any given deviation is positive or negative. We just want to know how far each observation is from the mean in any direction. So we need to transform the deviations into positive numbers that just measure the distance from the mean rather than the sign and distance. To do this, we could look at the absolute value of the deviations. But for reasons we'll discuss later, we typically make the deviations positive by squaring them instead. The variance is the average value of these squared deviations. So, if there are N observations (in our data, $N = 31$) the variance is

$$\sigma_X^2 = \frac{\sum_i^N (X_i - \mu_X)^2}{N}.$$

For the two variables in our data, the variances are

$$\sigma^2_{\text{crime}} = \frac{\sum_{i=1}^{31} (\text{crime}_i - \mu_{\text{crime}})^2}{31}$$

$$= \frac{(847 - 655.6)^2 + (555 - 655.6)^2 + \cdots + (708 - 655.6)^2}{31} \approx 5183$$

and

$$\sigma^2_{\text{temperature}} = \frac{\sum_{i=1}^{31} (\text{temperature}_i - \mu_{\text{temperature}})^2}{31}$$

$$= \frac{(-2.7 - 26.3)^2 + (-0.9 - 26.3)^2 + \cdots + (37.6 - 26.3)^2}{31} \approx 220.3.$$

By focusing on the average of the squared deviations rather than on the average of the absolute value of the deviations, the variance is putting more weight on observations that are farther from the mean. If the richest person in society gets richer, this increases the variance in wealth more than if a moderately rich person gets richer by the same amount. For example, suppose the average wealth is 1. If someone with a wealth of 10 gains 1 more unit of wealth, the variance increases by $\frac{10^2 - 9^2}{N} = \frac{19}{N}$. But if someone with a wealth of 100 gains one more unit of wealth, the variance increases by $\frac{100^2 - 99^2}{N} = \frac{199}{N}$.

The variance is a fine measure of how variable a variable is. But since we've squared everything, there is a sense in which it is not measured on the same scale as the variable itself. Sometimes we want a measure of variability that is on that same scale. When that is the case, we use the *standard deviation*, which is just the square root of the variance. We denote the standard deviation by σ (the lower-case Greek letter *sigma*):

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\frac{\sum_i^N (X_i - \mu_X)^2}{N}}.$$

The standard deviation—which is also a measure of how spread out a variable's distribution is—roughly corresponds to how far we expect observations to be from the mean, on average. Though, as we've noted, compared to the average absolute value of the deviations, it puts extra weight on observations that are farther from the mean.

For the two variables in our data, the standard deviations are

$$\sigma_{\text{crime}} = \sqrt{\frac{\sum_{i=1}^{31} (\text{crime}_i - \mu_{\text{crime}})^2}{31}}$$

$$= \sqrt{\frac{(847 - 655.6)^2 + (555 - 655.6)^2 + \cdots + (708 - 655.6)^2}{31}} \approx 72$$

and

$$\sigma_{\text{temperature}} = \sqrt{\frac{\sum_{i=1}^{31} (\text{temperature}_i - \mu_{\text{temperature}})^2}{31}}$$

$$= \sqrt{\frac{(-2.7 - 26.3)^2 + (-0.9 - 26.3)^2 + \cdots + (37.6 - 26.3)^2}{31}} \approx 15.1.$$

Now that we understand what a mean, variance, and standard deviation are, we can discuss three important ways in which we measure correlations: the *covariance*, the *correlation coefficient*, and the *slope of the regression line*.

Covariance

Suppose we have two variables, like crime and temperature, and we want to measure the correlation between them. One way to do this would be to calculate their *covariance* (denoted *cov*). To keep our notation simple, let's call those two variables X and Y. And let's assume we have a population of size N.

Here's how you calculate the covariance. For every observation, calculate the deviations—that is, how far the value of X is from the mean of X and how far the value of Y is from the mean of Y. Now, for each observation, multiply the two deviations together, so you have $(X_i - \mu_X)(Y_i - \mu_Y)$ for each observation i. Call this the *product of the deviations*. Finally, to find the covariance of X and Y, calculate the average value of this product:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

Let's see that the covariance is a measure of the correlation. Consider a particularly strong version of positive correlation: suppose whenever X is bigger than average $(X_i - \mu_X > 0)$, Y is also bigger than average $(Y_i - \mu_Y > 0)$, and whenever X is smaller than average $(X_i - \mu_X < 0)$, Y is also smaller than average $(Y_i - \mu_Y < 0)$. In this case, the product of the deviations will be positive for every observation—either both deviations will be positive, or both deviations will be negative. So the covariance will be positive, reflecting the positive correlation. Now consider a particularly strong version of negative correlation: suppose whenever X is bigger than average, Y is smaller than average, and whenever X is smaller than average, Y is bigger than average. In this case, the product of the deviations will be negative for every observation—one deviation is always negative and the other always positive. So the covariance will be negative, reflecting the negative correlation. Of course, neither of these extreme cases has to hold. But if a larger-than-average X usually goes with a larger-than-average Y, then the covariance will be positive, reflecting a positive correlation. If a larger-than-average X usually goes with a smaller-than-average Y, then the covariance will be negative, reflecting a negative correlation. And if the values of X and Y are unrelated to each other, the covariance will be zero, reflecting the fact that the variables are uncorrelated.

Correlation Coefficient

While the meaning of the sign of the covariance is clear, its magnitude can be a bit hard to interpret, since the product of the deviations depends on how variable the variables are. We can get a more easily interpretable statistic that still measures the correlation by accounting for the variance of the variables.

The *correlation coefficient* (denoted *corr*) is simply the covariance divided by the product of the standard deviations:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

When we divide the covariance by the product of the standard deviations, we are normalizing things. That is, the covariance could, in principle, take any value. But the correlation coefficient always takes a value between -1 and 1. A value of 0 still indicates no correlation. A value of 1 indicates a positive correlation and perfect linear dependence—that is, if you made a scatter plot of the two variables, you could draw a straight, upward-sloping line through all the points. A value of -1 indicates a negative correlation and perfect linear dependence. A value between 0 and 1 indicates positive correlation but not a perfect linear relationship. And a value between -1 and 0 indicates negative correlation but not a perfect linear relationship.

The correlation coefficient is sometimes denoted by the letter r. And we also sometimes square the correlation coefficient to compute a statistic called r-squared or r^2. This statistic always lies between 0 and 1.

One potentially attractive feature of the r^2 statistic is that it can be interpreted as a proportion. It's often interpreted as the proportion of the variation in Y *explained* by X or, equivalently, the proportion of X explained by Y. As we'll discuss in later chapters, the word *explained* can be misleading here. It doesn't mean that the variation in X causes the variation in Y or vice versa. It also doesn't account for the possibility that this observed correlation might have arisen by chance rather than reflecting a genuine phenomenon in the world.

Slope of the Regression Line

One potential concern with the correlation coefficient and the r^2 statistic is that they don't tell you anything about the substantive importance or size of the relationship between X and Y. Suppose our two variables of interest are crime and temperature in Chicago. A correlation coefficient of .8 tells us that there is a strong, positive relationship between the two variables, but it doesn't tell us what that relationship is. It could be that every degree of temperature corresponds with .1 extra crimes, or it could be that every degree of temperature corresponds with 100 extra crimes. Both are possible with a correlation coefficient of .8. But they mean very different things.

For this reason, we don't spend much time thinking about these ways of measuring correlation. We typically focus on the slope of a line of best fit, as we've already shown you. Moreover, we tend to focus on one particular way of defining which line fits best. Remember, a line of best fit minimizes how far the data points are from the line on average. We typically measure how far a data point is from the line with the square of the distance from the data point to the line (so every value is positive, just like with squaring deviations). We focus on the line of best fit that minimizes the sum of these squared distances (or the *sum of squared errors*). This particular line of best fit is called the ordinary least squares (OLS) regression line, and usually, when someone just says *regression line*, they mean *OLS regression line*. All the lines of best fit we drew earlier in this chapter were OLS regression lines.

The slope of the regression line, it turns out, can be calculated from the covariance and variance. The slope of the regression line (also sometimes called the *regression coefficient*) when Y is on the vertical axis and X is on the horizontal axis is

$$\frac{\text{cov}(X, Y)}{\sigma_X^2}.$$

This number tells us, descriptively, how much Y changes, on average, as X increases by one unit. Had we divided by σ_Y^2 instead of σ_X^2, then it would tell us how much X changes, on average, as Y increases by one unit. As we've seen, those can be different numbers.

We'll spend a lot more time on regression lines in chapters 5 and 10.

Populations and Samples

Before moving on, there is one last issue that is worth pausing to highlight. We can think about each of the statistics we've talked about—the mean, the variance, the covariance, the correlation coefficient, the slope of the regression line—in two ways. There

is a value of each of those statistics that corresponds to the whole population we are interested in. And there is a value of those statistics that corresponds to the sample of data we might happen to have. Either value can be of interest, but they can be importantly different. We have avoided that issue here by focusing on a case where our data and our population are the same—we have crime and temperature for every day in January 2018, which we've treated as our population and our sample. But this won't always be the case. For instance, we might have been interested in the relationship between crime and temperature in January over many years but only had a sample of data for the year 2018. This would give rise to all sorts of questions about what we can learn about January 2019 or January 1918 from our 2018 data. We will revisit these issues in chapter 6.

Straight Talk about Linearity

All of the various ways of measuring correlations that we have discussed focus on assessing linear relationships between variables. We will delve into this topic in more detail later on, especially in chapter 5 when we return to the topic of age and voter turnout in the context of our discussion of regression. But for now we will note that linear relationships are often interesting and important, but not all interesting and important relationships are linear. Consider, for example, the two possible relationships between the variables X and Y illustrated in figure 2.5.

As the regression lines make clear, in both these figures, the correlation between X and Y is 0. But these relationships are clearly different, just not in a way that is captured by the regression line.

In the left panel, there is no correlation between X and Y and there also doesn't seem to be any interesting relationship of any kind. You really can't predict the value of Y from X or vice versa. In the right panel, there is also no correlation between X and Y—on average, high values of X don't tend to occur with high values of Y, nor do low values of X tend to occur with low values of Y. But there is certainly a relationship between these two variables. In fact, X is quite useful in predicting Y in the right panel. This teaches us a lesson. Clear thinking about data requires more than just computing correlations. Among other things, it is important to look at your data (e.g., with scatter plots like these), lest you miss interesting nonlinear relationships.

There are lots of statistical approaches for dealing with non-linearity, and we'll discuss some of them in this book. But, as it turns out, linear tools for describing data can still be useful, even when the variables are related in a non-linear way. For instance, in the right panel of figure 2.5, there is a strong negative correlation between X and Y when X is less than 0 and a strong positive correlation between X and Y when X is greater than 0. So one thing we could do with linear tools is draw two lines of best fit, one for when X is less than 0 and one for when it is greater than 0. That would look like figure 2.6.

Another thing we could do is transform one of the variables so that the relationship looks more linear. For instance, in our example, although there is no correlation between Y and X, there is a strong linear relationship between Y and X^2. In figure 2.7 we plot X^2 on the horizontal axis and Y on the vertical axis. When we transform X into X^2, negative values of X become positive values of X^2 (e.g., -1 becomes 1), while the positive values stay positive (e.g., 1 stays 1). So it is as if we are folding the figure in

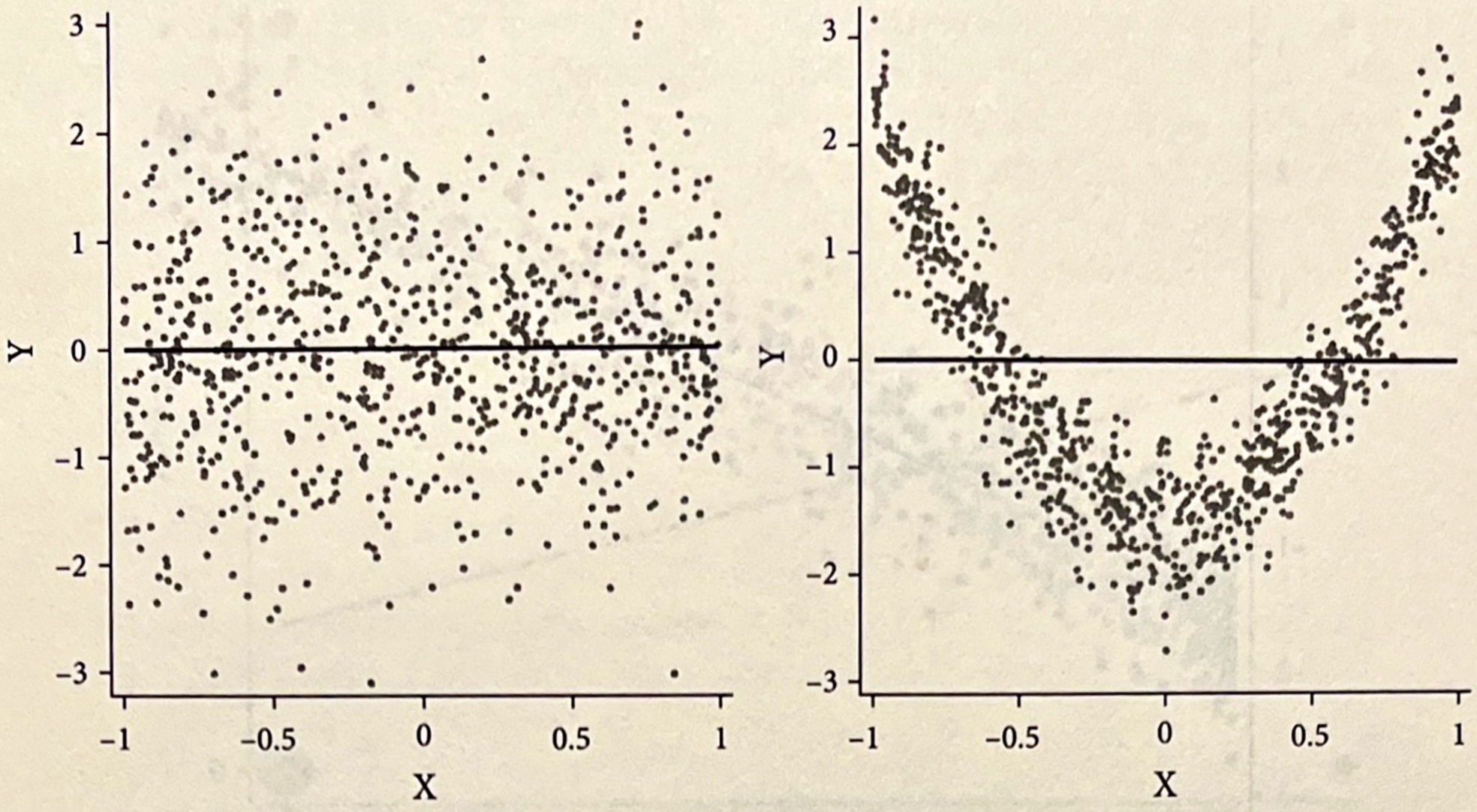

Figure 2.5. Zero correlation can mean many things.

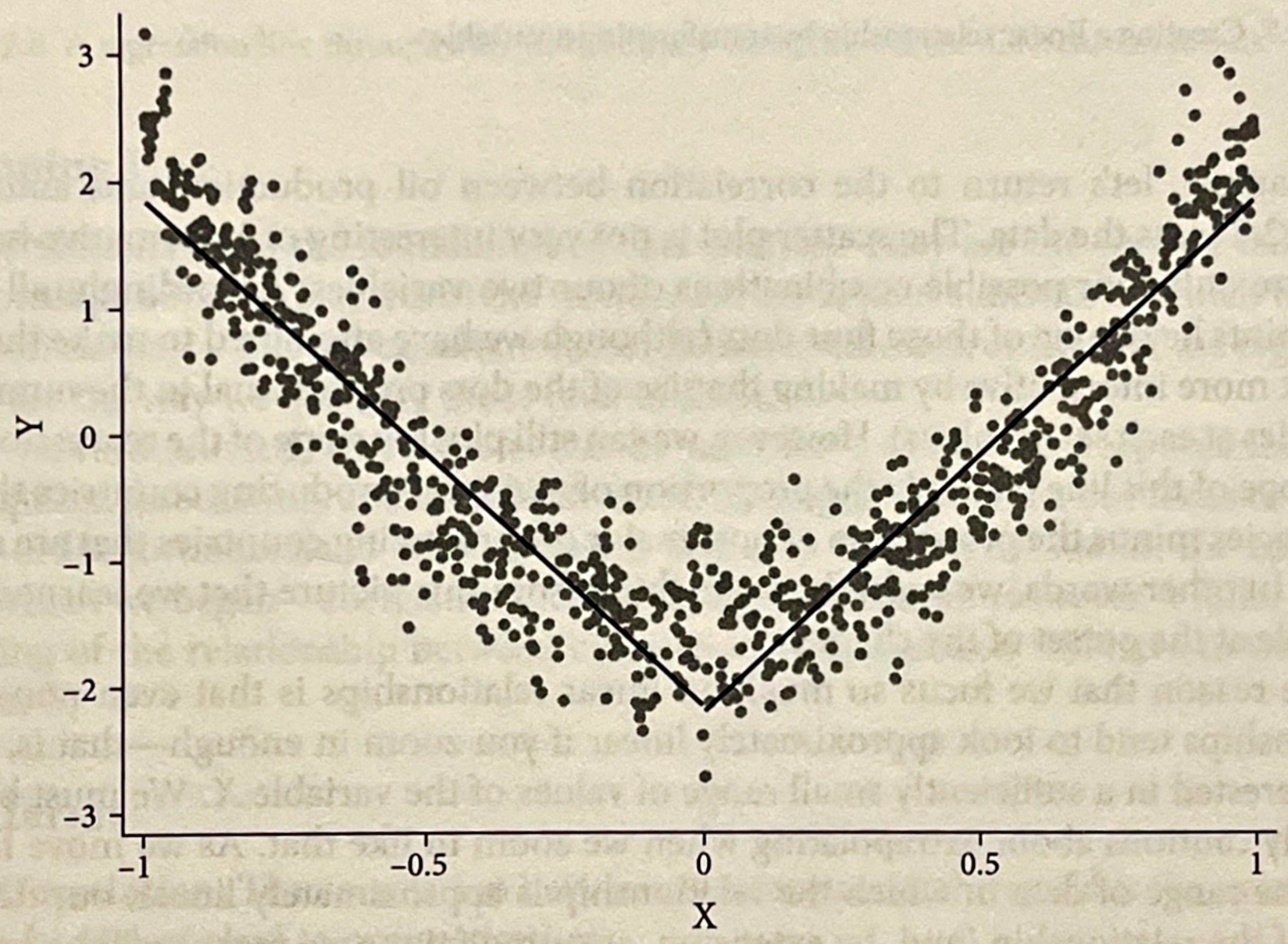

Figure 2.6. Fitting two separate regression lines to a non-linear relationship.

on itself at $X = 0$, and then we're twisting and stretching it a little so that X becomes X^2 (0 stays at 0, 1 stays at 1, .5 becomes $.5^2 = .25$, and so on).

With this transformation, our regression line shows a strong positive relationship between Y and X^2, and we can do a good job describing the relationship between these variables with our linear tools.

It's also worth pointing out that describing the relationship between two variables with a linear function is always appropriate when we're dealing with binary variables.

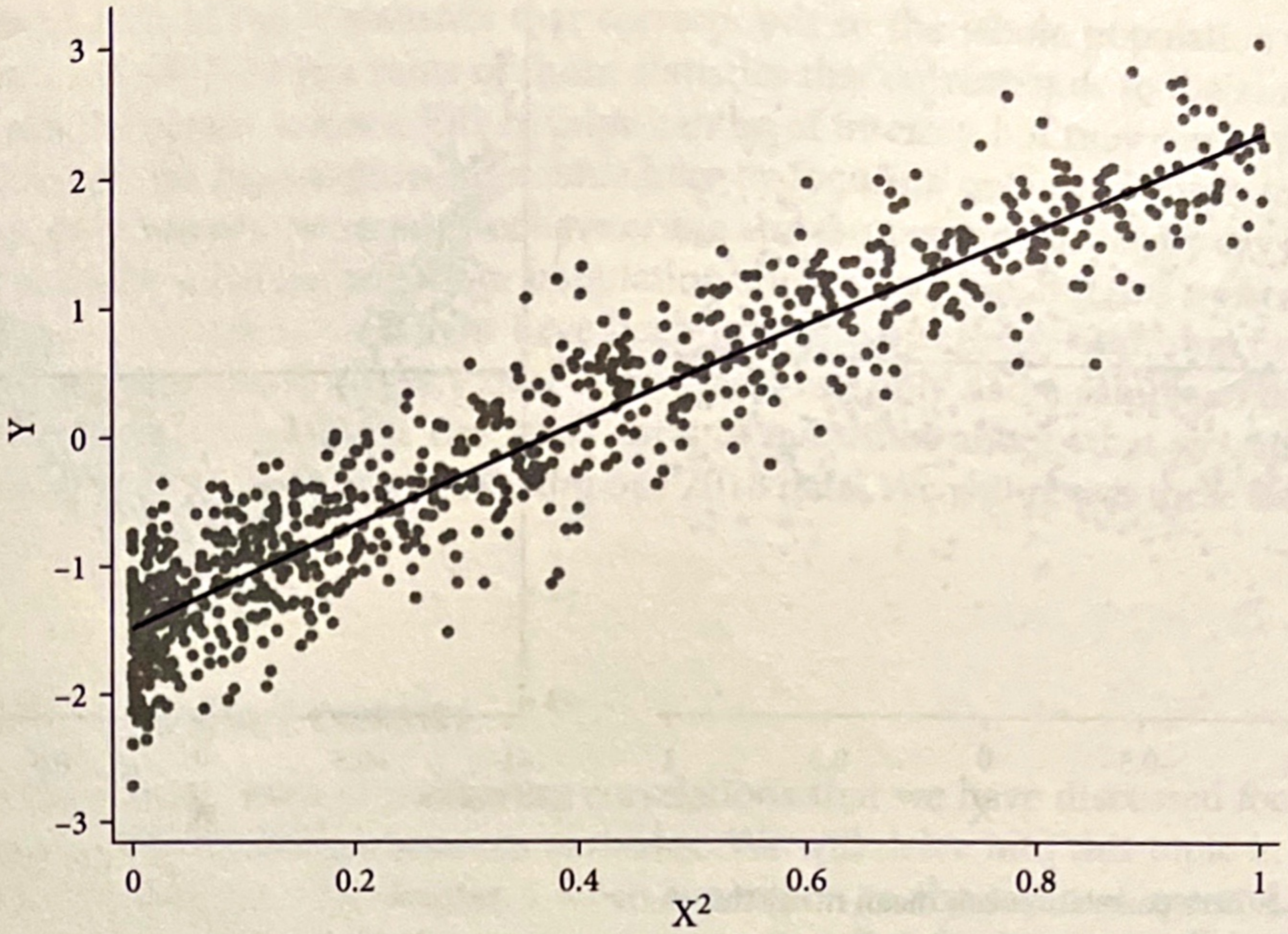

Figure 2.7. Creating a linear relationship by transforming a variable.

For example, let's return to the correlation between oil production and autocracy. Figure 2.8 plots the data. The scatter plot is not very interesting or informative because there are only four possible combinations of our two variables. Accordingly, all of the data points lie on one of those four dots (although we have attempted to make the scatter plot more informative by making the size of the dots proportional to the number of countries at each set of values). However, we can still plot the slope of the regression line. The slope of this line is simply the proportion of major oil-producing countries that are autocracies minus the proportion of non-major oil-producing countries that are autocracies. In other words, we learn the same thing from this picture that we learned from the table at the outset of the chapter.

One reason that we focus so much on linear relationships is that even non-linear relationships tend to look approximately linear if you zoom in enough—that is, if you are interested in a sufficiently small range of values of the variable *X*. We must be particularly cautious about extrapolating when we zoom in like that. As we move farther from the range of data in which the relationship is approximately linear, our descriptions of the relationship (and, by extension, any predictions we make) will be less and less accurate.

To think more about the dangers of extrapolation, consider an example. Political analysts find that the incumbent party in U.S. presidential elections tends to get about 46 percent of the vote when there is 0 income growth, and an extra 3.5 percentage points of the vote for every percentage point increase in income growth. Of course, they've measured this relationship using data on income growth levels that have actually occurred. Does this mean that we should predict incumbent vote share will be 81 percent if income growth is 10 percent? Probably not. And the incumbent's vote share definitely would not be 116 percent if income growth were 20 percent—that's impossible! But that doesn't mean a linear description of the data isn't useful for the range of income growths that we actually experience.

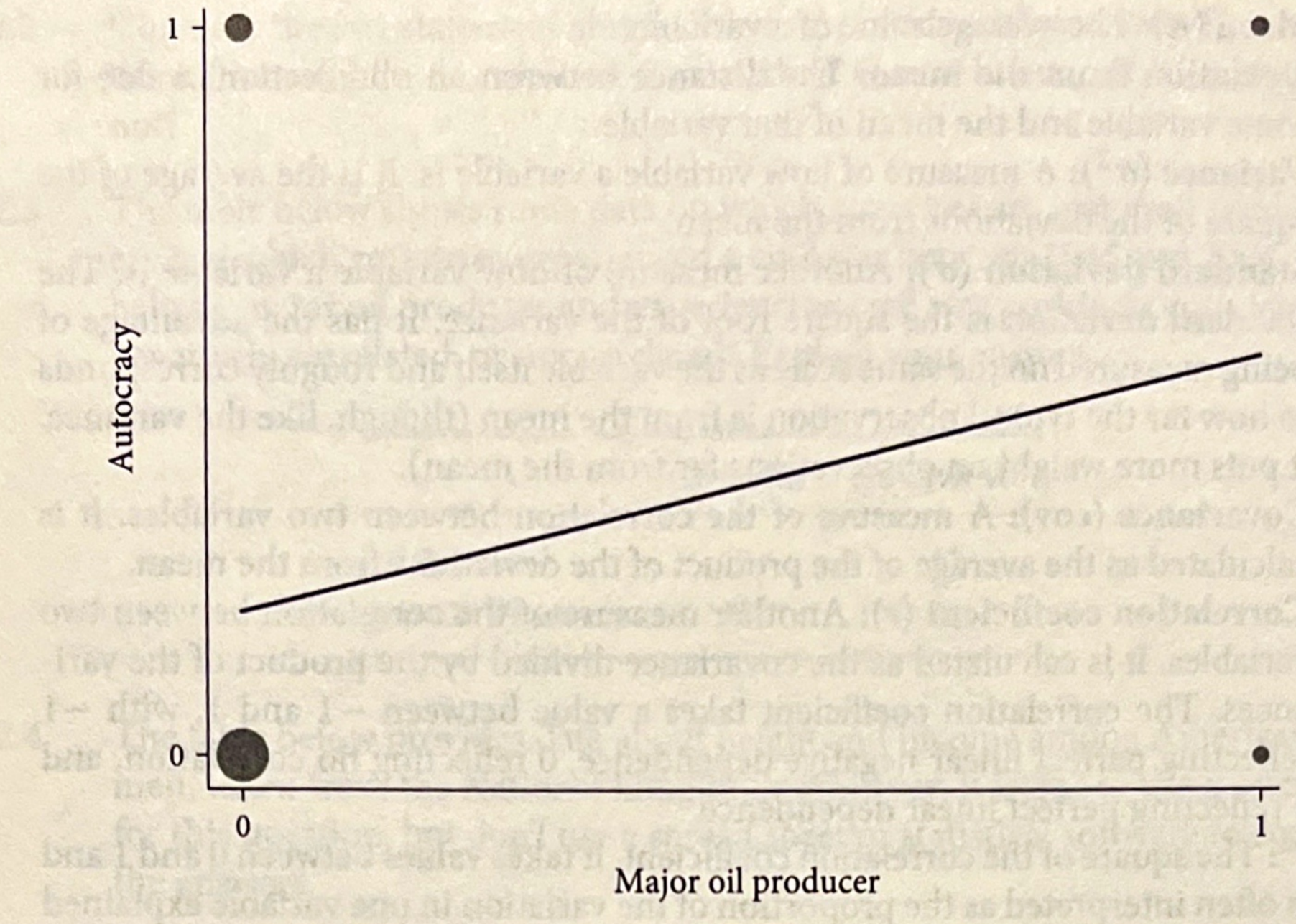

Figure 2.8. A regression line through data with a binary variable gives the difference in means.

Wrapping Up

Correlations form the foundation of data analysis. They are the way we talk about relationships between features of the world. And the various statistics by which we measure correlations—like the covariance, correlation coefficient, or slope of the regression line—are the way we quantify those relationships.

As we've discussed, correlations can be used for a variety of purposes including description, forecasting, and causal inference. In chapter 3, we turn our focus to causality in order to understand what it means and start to get a handle on the aphorism with which we began—correlation need not imply causation. However, a fuller understanding of the relationship between correlation and causation will have to wait until chapter 9.

Key Terms

- **Correlation:** The correlation between two features of the world is the extent to which they tend to occur together.
- **Positively correlated:** When higher (lower) values of one variable tend to occur with higher (lower) values of another variable, we say that the two variables are positively correlated.
- **Negatively correlated:** When higher (lower) values of one variable tend to occur with lower (higher) values of another variable, we say that the two variables are negatively correlated.
- **Uncorrelated:** When there is no correlation between two variables, meaning that higher (lower) values of one variable do not systematically coincide with higher or lower values of the other variable, we say that they are uncorrelated.
- **Line of best fit:** A line that minimizes how far data points are from the line on average, according to some measure of distance from data to the line.

- **Mean (μ):** The average value of a variable.
- **Deviation from the mean:** The distance between an observation's value for some variable and the mean of that variable.
- **Variance (σ^2):** A measure of how variable a variable is. It is the average of the square of the deviations from the mean.
- **Standard deviation (σ):** Another measure of how variable a variable is. The standard deviation is the square root of the variance. It has the advantage of being measured on the same scale as the variable itself and roughly corresponds to how far the typical observation is from the mean (though, like the variance, it puts more weight on observations far from the mean).
- **Covariance (cov):** A measure of the correlation between two variables. It is calculated as the average of the product of the deviations from the mean.
- **Correlation coefficient (r):** Another measure of the correlation between two variables. It is calculated as the covariance divided by the product of the variances. The correlation coefficient takes a value between -1 and 1, with -1 reflecting perfect linear negative dependence, 0 reflecting no correlation, and 1 reflecting perfect linear dependence.
- **r^2:** The square of the correlation coefficient. It takes values between 0 and 1 and is often interpreted as the proportion of the variation in one variable explained by the other variable. But we have to pay careful attention to what we mean by "explained." Importantly, it doesn't mean that variation in one variable causes variation in the other.
- **Sum of squared errors:** The sum of the square of the distance from each data point to a given line of best fit. This gives us one way of measuring how well the line fits/describes/explains the data.
- **OLS regression line:** The line that best fits the data, where *best fits* means that it minimizes the sum of squared error.
- **Slope of a line:** The slope of a line tells you how much the line changes on the vertical axis as you move one unit along the horizontal axis. So a completely horizontal line has a slope of 0. An upward sloping 45-degree line has a slope 1, a downward sloping 45-degree line has a slope of -1, and so on.
- **Slope of the regression line or regression coefficent:** The slope of the regression line describes how the value of one variable changes, on average, when the other variable changes. The slope of the regression line is the covariance of two variables divided by the variance of one of them, sometimes also called the regression coefficient.

Exercises

2.1 Consider the following three statements. Which ones describe a correlation, and which ones do not? Why?

(a) Most professional data analysts took a statistics course in college.
(b) Among Major League Baseball players, pitchers tend to have lower-than-average batting averages. (We'll learn why this is the case in chapter 16.)
(c) Whichever presidential candidate wins Ohio tends to win the Electoral College.