1

Introduction

This chapter provides background for the multiple regression methods that are presented in the remainder of the book. First, we discuss some of the potential advantages of quantitative research in the social sciences, along with some important cautions to heed. Second, we review some core statistical ideas, including descriptive statistics, hypothesis tests, correlation, and simple (bivariate) regression, which together form a foundation for understanding multiple regression. We illustrate those ideas with a full example and an exercise. Although we expect readers to have been exposed to these statistical concepts before, the material in this chapter will be a helpful refresher.

1.1 Quantitative Research

Social scientists engage in various kinds of research using different research methods, and the quantitative approach explored in this book does not characterize all of social science. However, this approach has proven helpful in investigating a wide variety of questions in social science and has been the main mode of inquiry in fields such as sociology, psychology, criminology, political science, and public health for many years. Before diving into the statistical material that makes up the remainder of this chapter, it is useful to discuss in brief some of the main attributes that make quantitative research approaches attractive as tools for social science research. Although this list is not exhaustive, we hope that it gives readers a sense of why the effort to learn these methods is worthwhile. We will also provide some cautions, as it is important to understand the limitations of these tools along with their strengths.

1.1.1 Benefits of Quantitative Research

One of the most appealing features of quantitative research is its openness to the scrutiny of others. When we disseminate a report or article describing a research study that we have conducted, critics can examine all parts of our work, including the decisions we made in preparing our data for analysis, our choices of analytic methods, and the interpretations we drew from our analytic results. They can attempt to verify our findings by repeating our analysis and confirming that we made no errors in applying the analytic methods that we used, or argue for alternative analyses and see if those lead to the same scientific conclusions about our research question. Critics can even look further back in the research process and challenge the methods that we, or others, used to collect data, or take issue with the overall design of the study. Even if this scrutiny does not suggest any problems with our study in isolation, other researchers can also check if our findings hold up when they analyze new data.

All of this can be uncomfortable for the researchers whose work is being examined; naturally no one enjoys being criticized, and it is hard not to take academic criticism personally. But this scrutiny is an essential part of an academic discipline's journey toward understanding, and a field's progress is held back if conclusions from flawed research are accepted as accurate. Note that this is not simply an issue for academic debates, because there can also be important consequences in the larger world when research is used to craft policy initiatives. Of course it is not always easy in practice for a field to carry out intensive scrutiny of the research that it produces. Researchers may resist requests to share their data with others, or research reports may not be clear enough for others to be sure as to precisely what methods were used to collect or analyze data. Still, even with these practical challenges, quantitative research is more amenable to this scrutiny than are other types of research that do not rely on formalized procedures for collecting and analyzing data, or for which researchers do not have a well-defined data set that can be shared with others. Although there surely is also critical examination of academic work that relies on other kinds of research approaches, the nature of quantitative research makes this scrutiny especially possible and productive.

Along with facilitating critical examination of our research, quantitative methods can act as a check on ourselves when drawing conclusions from our research. Even if our intent is to be completely accurate and fair in our analysis and interpretations, human nature sometimes gets in the way. People are inclined to perceive patterns in data even when there really are not any, or put greater emphasis on observations that support what they already believe than on observations that contradict their beliefs. Although conscious misrepresentation of data or analytic results can occur, cases of that sort of intentional fraud or deception are likely rare. The bigger concern is that researchers who are honestly trying to be accurate and fair are still subject to these human flaws and biases.

Quantitative research, with its formal procedures for analyzing data and interpreting results, helps guard against these dangers. Patterns and relationships in data are deduced from formal analysis, rather than informal synthesis of observations, and researchers need to justify any decisions to exclude observations from

their quantitative analysis. Of course researchers still need to make a wide variety of decisions when collecting data and choosing analytic methods, and some choices could potentially serve to push final results in a direction desired by the researchers. Likewise, interpretations of results are not always straightforward even when standard analytic procedures are used and therefore are subject to these dangers as well. So these threats cannot be completely eliminated even in carefully executed quantitative research. Nonetheless, quantitative research goes much further in addressing these concerns than research that essentially relies on the unaided human mind to carry out analysis and interpretation. We should be skeptical of such research no matter how intelligent the researchers are.

Another beneficial aspect of quantitative research is that quantitative measurement forces us to be as precise as possible about the social scientific concept that we intend to measure and how we can or cannot effectively measure it. Of course in any kind of empirical research we need to think carefully about the logic of research design, and all research likewise requires us to link broad concepts to the actual observations that are used in our study. Still, this linkage is especially salient in quantitative research, as we directly confront the issue of expressing theoretical concepts through concrete measurements. Some quantitative studies fall short in this regard, and the failure of a measure to adequately capture its intended concept is often a point of criticism when research projects are evaluated. But even if a measure is lacking in this way, the criticism will point to improvements that can be incorporated in further research. Formal measurement is therefore a key mechanism through which researchers can collectively work toward a better understanding of the world. When measurement is informal, it is much more difficult to even begin that conversation.

In practice, a good deal of social science research is done with at least some intent that the research will contribute to the development of effective public policy around some social problem or goal. The numerical findings that summarize the results of quantitative research can be especially helpful in this respect, as they can facilitate cost/benefit calculations that attempt to weigh the potential beneficial impact of a policy initiative against the expense required to implement it. For example, quantitative studies could help estimate the expected impact that a given increase in a school district's number of teachers may have on its high school graduation rate. Community members could then use this estimate as one element in deciding whether it will be worthwhile to direct more funds toward hiring teachers and, by implication, away from other programs supported by those funds. Even with quantitative estimates available, the ultimate decision will of course still require the school district to assess its priorities and make hard decisions about the services that it can realistically offer in light of its fiscal constraints, so the decision to add teachers surely involves much more than a single cost/benefit calculation. But relevant research results can be part of the debate in a much more helpful way than would be possible if the research findings could not be expressed numerically.

Another benefit of quantitative research is that it positions social scientists to take advantage of the ongoing development of new methods in the field of statistics. Even though there are important examples through the years of quantitative methods that were originally developed by social scientists,

a substantial portion of the analytic methods that practicing social scientists use every day actually began in other academic fields, especially statistics. Statistical research focuses on developing and refining such methods and on understanding the conditions under which methods do or do not work as expected. By importing these methods and interpretations when required, social scientists can instead focus their intellectual energy on the formulation of social scientific research questions and how to add to our knowledge of the social world and improve our theories about it. Naturally a book like ours focuses on "workhorse" methods that have come to be the most widely used by social scientists over many years, rather than new methods that are at the cutting edge of statistical development. But the quantitative approach to research facilitates the eventual absorption of what are now cutting-edge methods into standard social scientific practice.

Finally, quantitative research is fun. Students often find that learning quantitative methods feels like coming upon an entirely new set of tools that can, if used appropriately, yield tremendous insight into many of the most important questions in social science. We hope that readers of this book will share this feeling of excitement.

1.1.2 Some Cautions

Quantitative methods offer tremendous promise, but researchers also need to be cautious in applying the methods and interpreting results. Some of these cautions will be mentioned in the context of specific methods introduced in later chapters, but it is helpful to highlight a few general issues here.

First, it is important to be clear that statistical analysis of quantitative data, even if highly sophisticated, technically advanced, and executed by knowledgeable researchers, typically cannot compensate for fundamental problems in the research design and data collection efforts underlying the data set being analyzed. In this book, we do not give much attention to all the parts of the research process that precede data analysis. But that is due only to limits of scope and space; careful attention to research design and data collection is absolutely crucial to carrying out successful and meaningful research. There are limited circumstances in which analytic methods have some ability to compensate for "bad" data after they have been collected, but such circumstances should be viewed as the exception, not the rule. The old adage "garbage in, garbage out" is usually pretty accurate. Although we focus here on quantitative analysis, in no way do we intend to minimize the vital importance of appropriate research design and data collection as discussed in social science research methods texts and courses.

A broad theme is that we need to guard against overinterpreting our research findings, or believing that they tell us more, and more definitively, than they really do. This general warning can take many specific forms. For instance, we must recognize that there will always be some uncertainty in the results that we report, at least if we intend to somehow generalize our findings beyond description of the particular sample used in our analysis. Some aspects of uncertainty will be addressed by statistics that we can report in our analysis, such as standard errors or hypothesis tests. But there is also an important sense in which a

single study can rarely be truly conclusive, and a finding needs to be replicated in data from other samples or other contexts before we can begin to really trust it. Literal replication can be difficult for much research in the social sciences. Research may be situated in a historical setting that cannot be revisited, or in a *population* from which repeated sampling is unrealistic. But even if identical conditions cannot be repeated, and literal replication is impossible, we still will find research results more compelling if they seem to hold up in similar settings or populations. It is also crucial to acknowledge uncertainty in research findings when communicating with policymakers.

Another kind of overinterpretation is the *ecological fallacy*. As typically discussed in research methods texts and courses, the ecological fallacy refers to the distinction between the aggregate and individual *level of analysis*. The aggregate level refers to units like cities, states, or nations, representing collections of individuals and for which we can collect aggregated information like the percentage of the population that is unemployed or the average years of education for its adults. Individual-level units are, as the name suggests, the individuals themselves rather than collections of them. Then we could collect information about each individual, such as whether or not an individual is unemployed or the number of years of education that a person has completed. Distinctions between levels sometimes are not obvious. For instance, for some purposes a multiperson household may seem like an aggregate-level unit composed of a set of individuals, but for other purposes we envision a household making decisions and otherwise behaving as if it were an individual unit. For now, we will ignore this and suppose that there is a clear line between individual- and aggregate-level units.

On one hand, it is often easier for a researcher to obtain data on aggregate-level rather than individual-level units. Although many of the statistics that are reported for cities, counties, or states are ultimately based on surveys of individuals, or collected from official records that governments maintain on individuals, the aggregate-level information may be distributed in a way that makes it more accessible to researchers than the original individual-level survey data or records. And while data from many individual-level surveys are publicly available, a researcher may find that none of them include all of the desired information. (Because the same individuals generally do not appear in different surveys, it is usually not feasible to obtain data from more than one survey.) Sometimes individual-level measures can be obtained only by launching a new survey, but often that would be quite difficult and expensive. On the other hand, many of the theories and research questions that most interest social scientists actually lie at the individual level. For instance, the individual-level question of whether a person's employment status predicts their criminal behavior may seem more interesting than the aggregate-level analogue asking whether a city's unemployment rate predicts its crime rate.

The main objective is for the level of analysis implicit (or explicit) in the research question being studied to match the level of analysis at which the data used to address the research question were actually collected. That is, an individual-level research question should be considered in light of individual-level data. But noting the often greater accessibility of aggregate-level data alongside the attractiveness of individual-level research questions, it is not surprising that

there is sometimes a mismatch. (The mismatch can be in the opposite direction too, with individual-level data being used to address an aggregate-level research question, but that is a less common problem.)

The ecological fallacy, then, refers to attempts to answer individual-level research questions via aggregate-level data. Using the example above, a finding of a relationship between city-level unemployment and crime rates cannot be taken as evidence that unemployed people are more likely to commit crime. It could be instead that within each city there is no pattern of unemployed people committing disproportionately many crimes, but rather that in places with higher unemployment, more crime is committed by *both* employed and unemployed people. Even if aggregate-level data should not be used to address individual-level questions or draw individual-level conclusions, researchers may find it very hard to resist doing so. The ecological fallacy is therefore a specific instance of the more general threat of overinterpretation of analytic results.

Naturally these few cautions do not exhaust the threats that one must be aware of when applying quantitative research methods. As we explore various specific techniques in the chapters that follow, we will sometimes warn of possible misinterpretations or common errors. These warnings should not dampen the excitement of working with quantitative research, but they do remind us that an analyst must always strive to exercise good judgment in using these tools.

1.2 Review of Basic Statistics

We conclude this chapter by reviewing some of the core ideas that are taught in typical beginning statistics courses for students of social science. We assume that readers have already been exposed to this material through a formal course or self-study. This brief review thus will be a refresher that leads into the new material covered in the rest of the book.

1.2.1 Descriptive Statistics

Descriptive statistics summarize the values of the variables for which information on the cases (the units of analysis) in a sample is available. The simplest description of a variable is the frequencies of the different values it takes in the sample. A *frequency distribution* or table typically reports the number and percentage (or relative frequency) of sample cases that take on each observed value of the variable.

Frequencies will be most appropriate for use with *categorical variables* (such as political affiliation) in which the values indicate category membership, or with discrete numerical variables for which there is a limited range of observed numerical values (for instance, the number of children that a person has). Frequencies can be impractical for use with continuous numerical variables (like height), which may take on very many different values among the sample's cases. Such a variable could even have a unique value for every case in the sample, especially if it is measured with great precision.

For a numerical variable that was measured on the sample cases, we can instead start by reporting the *mean* (usually denoted by a bar over the variable's

label, such as $\overline{X}$) value of that variable. The mean is just a different name for the commonsense idea of an average: we simply add up the variable's values across all of the cases and divide by the number of cases N. The mean is interpreted as the typical value of X in the sample, so it is sometimes called a measure of central tendency. The *median* is an alternative measure of central tendency and is defined as the middle value after rewriting the sample values of X in order from lowest to highest. (If N is an odd number, it is clear which value is in the middle of the list; if N is even, one could take the average of the two middle values.) The median is especially appealing when the sample includes a small number of very high or low values that might distort the mean, but the mean is more important for the statistics that we discuss in this book.

Along with describing the typical value of a variable in a sample, it is important to also know how spread out the variable's values are. One way to measure this is the *range*, or simply the minimum and maximum values that the variable takes across the cases in the sample, along with the difference between those. The range is straightforward and informative, but one or two extreme values can of course have a big impact on it. The *variance* (usually denoted by s^2) is a more comprehensive measure of spread. It is based on the squared deviation from the mean of each case's value on the variable; these squared deviations are then summed over all cases and the sum is divided by N or (N – 1). For interpretation, the variance is usually transformed into the *standard deviation* (denoted by s) by taking its square root. The standard deviation can be interpreted as roughly the typical absolute (that is, ignoring positive and negative signs) deviation from the mean of the variable for cases in the sample.

1.2.2 Hypothesis Tests

In classical hypothesis testing, the analyst attempts to decide which of two distinct propositions, the *null hypothesis* (denoted H_0) and the *alternative hypothesis* (H_1), about some aspect of the population seems more plausible in light of the sample data. In this framework, the null hypothesis expresses current knowledge or belief about that population characteristic, while the alternative hypothesis is thought to hold if the null is not correct. (Note that in practice, the null may just be a convenient starting point for the research, rather than actually reflecting genuine knowledge or belief.) The analyst uses the sample data to decide whether or not to reject the null, where rejecting the null means that the alternative was found to be more plausible. One might see failure to reject the null as logically equivalent to "accepting" it, but researchers usually avoid that language; in principle, future research could lead to rejection of the null even if that did not occur in this study, so "accepting" seems too strong.

When the null indicates a particular value for some population characteristic, the *hypothesis test* is based on a comparison of the hypothesized value to its estimate in the sample. For example, if the null hypothesis is that the mean age in a population is 41 years, the mean age in the sample is compared to the hypothesized value 41. This comparison takes into account some assessment of sampling variability in the sample statistic; in the case of a single mean, the *standard error of the mean* represents sampling variability and is typically

estimated by $s/\sqrt{N}$. A test statistic formalizes the comparison, using the form $\frac{\text{(Sample value – Hypothesized value)}}{\text{Standard error}}$. For a single mean, with the hypothesized population mean denoted by μ_0, this test statistic becomes $(\overline{X} - \mu_0) / \left(s/\sqrt{N} \right)$.

If the null hypothesis is correct, statistical theory will indicate a particular probability distribution for the test statistic (in our example of a single mean, this will be a t-distribution). This allows calculation of a *p-value*, defined somewhat loosely as the probability, under the assumption that the null is correct, of obtaining a sample value that is as discrepant (or more) from the hypothesized value as was the statistic that we calculated in our sample. In the example of a single mean, suppose that $\mu_0 = 41$ and our sample yielded a mean $\overline{X} = 45$. Then the p-value, calculated by assuming that the population mean really is 41, reports the probability of obtaining a sample mean as far (or farther) from 41 as was our sample mean 45.

A small p-value pushes us toward rejecting H_0, because it suggests that either (a) the null really is true, but we happened by chance to get quite an unusual sample, or (b) our estimate in the sample just seems unusual because we calculated its probability under an incorrect premise in H_0. If the p-value is small enough, we will follow the logic of (b) and reject H_0, but we do that with a recognition that (a) could actually be what happened. The cutoff for deciding what is a "small" p-value is a statement of how concerned we are about incorrectly rejecting H_0 under scenario (a). The more worried we are about that possibility, the smaller the cutoff we would use, because a smaller cutoff leaves us less vulnerable to incorrectly rejecting H_0. In practice, there is a long tradition of using 0.05 as the cutoff for a "small" p-value, so that we reject H_0 if $p \leq 0.05$, and do not reject if $p > 0.05$. But it is important to recognize that the choice of 0.05 is essentially arbitrary, even though there is a long history of its use.

When H_0 indicates a particular value for the population (like $\mu_0 = 41$ in our example), we need to note whether the alternative is *one-sided* or *two-sided* (also known as one-tailed or two-tailed). This distinction refers to whether the alternative hypothesis proposes values that are only on one side of the hypothesized value, or on both sides. In our example of a single mean, a one-sided H_1 would indicate $\mu > 41$, or perhaps $\mu < 41$; either way, the alternative states that conceivable values of μ lie only above, or only below, the hypothesized value. A two-sided alternative accommodates both of these possibilities; in our example, this would be $\mu \neq 41$. The choice of a one-sided or two-sided H_1 does make a difference when calculating the p-value. The two-sided p-value is twice the one-sided p-value, because in the two-sided case we are asking about the probability that a sample statistic will be so discrepant from the hypothesized value in either direction. If we have enough understanding of the research topic to absolutely rule out the possibility of population values on one side of the hypothesized value, then it is good to use the one-sided alternative. But often we do not have such strong understanding that we can rule out these values, and the two-sided alternative would be appropriate.

Continuing with our example, suppose that we set up a two-sided alternative hypothesis, so that the null and alternative hypotheses were

- H_0: $\mu = 41$
- H_1: $\mu \neq 41$.

Suppose also that along with calculating $\overline{X} = 45$ from the data in a sample of 125 individuals, we calculated the standard error of $\overline{X}$ ($s/\sqrt{N}$ from above) to be 2.62. Then the test statistic is (45 – 41) / 2.62, or 1.53. We obtain the tail probability associated with the value 1.53 for a t-distribution with (125 – 1) = 124 *degrees of freedom* (df) from a table or online calculator. When using a table or online calculator, we need to note whether the alternative hypothesis is one-sided or two-sided. If we are using an online calculator that provides one-sided probabilities only, t = 1.53 corresponds to probability 0.064. Because H_1 was two-sided, we need to use both the upper and lower tails, which doubles the probability. Our p-value is therefore $0.064 \times 2 = 0.128$. However, we would obtain the value 0.128 directly if the online calculator provides two-sided probabilities. As this p-value is > 0.05, we do not reject H_0. Our conclusion is that the sample data are not so discrepant from the null hypothesis as to convince us that the null is implausible, and for now we believe that the hypothesis that the population's mean age is 41 remains reasonable.

When the null hypothesis is more complicated than simply indicating the value of a single population characteristic, the development of a test statistic and in turn the determination of a p-value will be more involved. But even if getting to the p-value is more complicated, the final step of comparing it to a cutoff (usually 0.05) to decide whether or not to reject H_0 will still apply. The chapters to follow will include various tests of hypotheses that are relevant in multiple regression analysis.

1.2.3 Correlation

Pearson's *correlation* or correlation coefficient (r_{xy}) allows us to assess the direction and strength (as defined below) of a relationship in our sample data between two numerical variables X and Y. (We do not explore any other types of correlation in this book, so for convenience we will drop the name "Pearson's" when discussing r_{xy}.) The direction of the relationship can be positive or negative, or there could be no relationship at all. In a positive relationship, the general tendency is for cases that have a high value of X to also have a high value of Y, and, conversely, for cases that have a low value of X to also have a low value of Y. High values of the two variables tend to be seen together, and likewise for low values. Figure 1.1 shows a scatterplot of values of X and Y for a hypothetical sample with 12 cases: a point is plotted at the (X, Y) values for the case it represents.

FIGURE 1.1 ● A Scatterplot of Values of X and Y (Positive Relationship)

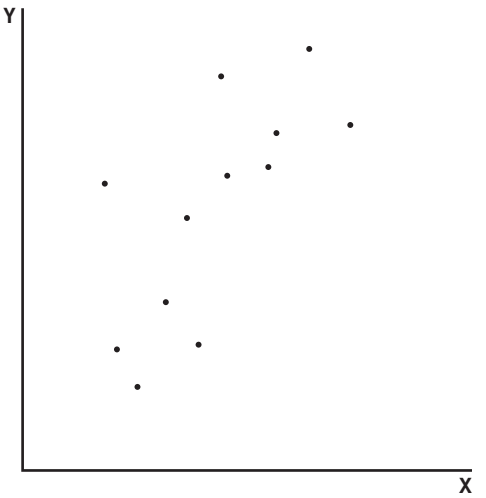


Figure 1.1 illustrates a positive relationship. Cases that have high values of X—toward the right side of the plot—also tend to have high values of Y, and cases with low values of X—toward the left of the plot—tend to have low values of Y. The cloud of points has a general shape that goes from lower left to upper right in the plot.

In a negative relationship, cases with low values of X tend to have high values of Y, and cases with high values of X tend to have low values of Y. That is, high and low values tend to be opposite for the two variables across cases. Figure 1.2 illustrates a negative relationship.

In Figure 1.2, cases on the left, with low values of X, tend to have high values of Y, and cases on the right, with high values of X, tend to have low values of Y. The cloud of points has a general shape of upper left to lower right. Compared with Figure 1.1, however, the pattern seems a bit less distinct here. For instance, the cases with the lowest values of Y in the figure are only average or slightly above average on X, rather than having especially high values of X. Still, the overall pattern is that of a negative relationship.

When there is no relationship between X and Y, the cloud of points does not exhibit either of the general shapes that indicates a positive or negative relationship. There is no tendency for cases with high (or low) values of X to have either high or low values of Y. Figure 1.3 shows a sample in which there is no apparent relationship between X and Y when we consider all points jointly.

As a result of the method by which the correlation coefficient r_{xy} is calculated, its value can range between –1 and 1. Its sign—positive or negative—indicates the direction of the relationship between X and Y. A positive value of r_{xy} corresponds to a positive relationship, and a negative value of r_{xy} corresponds to a negative relationship. If there is no relationship between X and Y, the value of r_{xy} will be zero.

FIGURE 1.2 ● A Scatterplot of Values of X and Y (Negative Relationship)

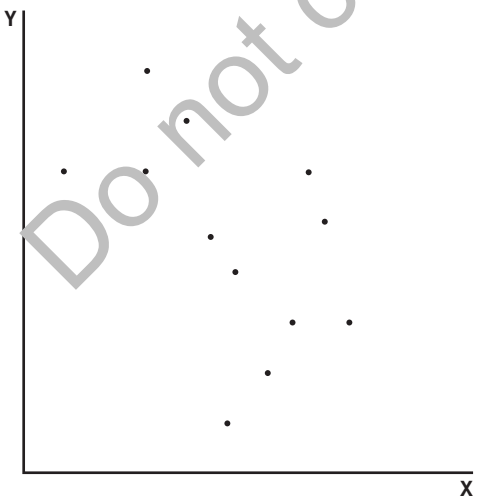
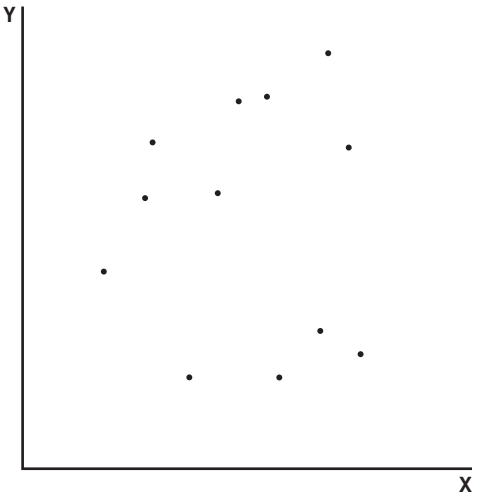


FIGURE 1.3 ● A Scatterplot of Values of X and Y (No Relationship)



The magnitude of r_{xy} is also informative. It indicates the strength of the relationship between X and Y, in which a stronger relationship means that the points in the scatterplot are closer to lying on a single straight line. If the points are close to lying on a straight line, then for any particular value of X there is a very narrow range of Y values that appears along with it in the sample (and vice versa). In that situation, knowing what value a case had on one of the variables will give a very good idea of the likely values of the other variable. But if the points are far from lying on a straight line, then knowing the value of one of the variables is not so informative about the likely value of the other variable. In that situation the variables have a weaker relationship.

As mentioned above, r_{xy} can be no bigger than 1 and no smaller (more negative) than –1. Either of these extreme values indicates a perfect relationship between X and Y in the sample, with the points in the scatterplot exactly forming a straight line. (The perfect relationship is negative or positive depending on the sign of r_{xy}.) With a perfect correlation, one can precisely determine a case's value of one variable from its value on the other.

As r_{xy} gets closer to 0, indicating a weaker relationship between X and Y in the sample, the cloud of points looks less and less like a straight line. The sign still shows the direction of the relationship, but values close to 0 indicate a weak relationship, in which a case's value on one variable is not very informative about its value on the other. Figure 1.2 shows a somewhat less distinct relationship than Figure 1.1. That is borne out by the correlation coefficients for the data in the two figures. The data of Figure 1.1 result in $r_{xy} = 0.70$, suggesting a fairly strong positive relationship between X and Y. However, the data of Figure 1.2 result in $r_{xy} = -0.50$, a more moderate negative relationship. The points in Figure 1.1 are therefore closer to lying on a straight line than those in Figure 1.2. (The data of Figure 1.3 result in $r_{xy} = 0.00$, corresponding to no relationship.)

It is important to note that when we discuss the relationship between X and Y in terms of r_{xy}, we are actually referring only to a straight line, or *linear relationship*. It is possible for a scatterplot to show a discernible pattern of how X and Y values go together in the sample but still give an r_{xy} close to 0 because that pattern is not consistent with a straight line. Usually a low r_{xy} genuinely does indicate no relationship between X and Y, but we should keep in mind that a more complicated *nonlinear relationship* could at least theoretically be present even when r_{xy} is close to 0.

1.2.4 Linear Regression

We can describe the relationship between X and Y further by performing a *simple (*or *bivariate) linear regression*. In the correlation, X and Y played equivalent roles, so we could haphazardly label one variable as X and the other as Y and it would not make a difference. But when we move to regression, we need to designate one variable as the dependent variable and one as the independent variable. The dependent variable is the outcome that we are attempting to explain or predict, while the independent variable is the factor that we believe predicts or, to be a bit bolder, determines the value of the dependent variable. The convention is that the dependent variable is labeled Y and the independent

variable is labeled X. "Simple" linear regression refers to this situation of a single independent variable X influencing Y, but linear regression can be extended to include more than one independent variable, and that will be the focus of the rest of the book. For the moment, however, we will consider only analyses involving a single X. Our setup for the (simple) linear regression therefore can express a primitive theory in which X influences Y, even though we will often recognize that our research design does not allow us to convincingly say that the value of X truly "causes" the value of Y.

Linear regression fits a straight line (the "*regression line*") to the cloud of points in the scatterplot. This line is the best (in a certain sense, to be discussed below) possible straight-line summary of the relationship between X and Y that is observed in the sample. In the discussion in Section 1.2.3, we saw that when r_{xy} is closer to 1 or –1, the summary provided by the line becomes a more accurate description of how X relates to Y in the sample data. That is because a high correlation indicates that the cloud of points is itself closer to being a straight line. As we mentioned with the correlation, the relationship between X and Y could be more complicated than can be represented by a straight line. Although linear regression inherently focuses on straight-line relationships, in later chapters we will see how the regression framework can actually accommodate various more complicated situations too.

Figures 1.4 and 1.5 show the same plots of 12 points displayed in Figures 1.1 and 1.2, but with the regression line added to each. In Figure 1.4, we see that the line has a positive *slope* (lower left to upper right), which agrees with our informal inspection of the cloud of points before, and is consistent with the sign of the correlation that we reported earlier. In Figure 1.5, the line has a negative slope (upper left to lower right).

FIGURE 1.4 ● Regression Line With a Positive Slope

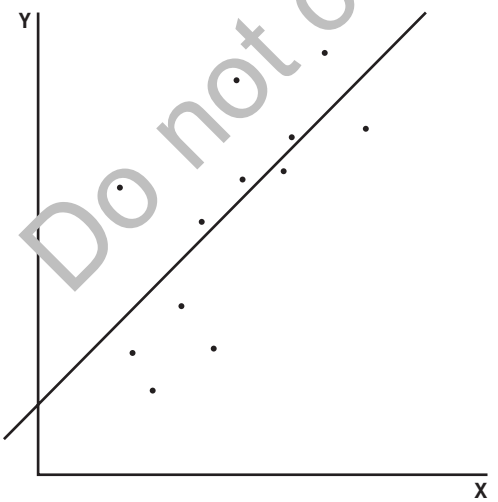
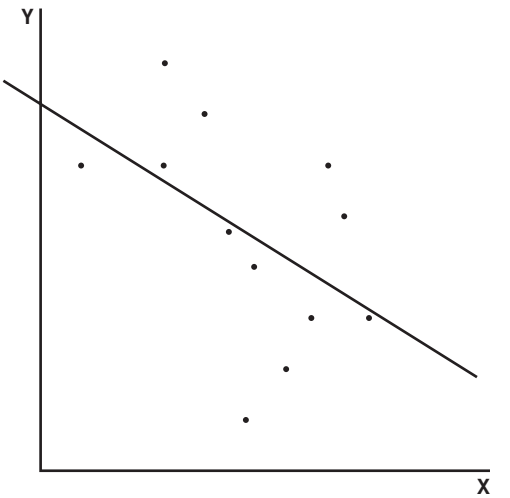


FIGURE 1.5 ● Regression Line With a Negative Slope



We can think of the line as representing a prediction of the value of Y for any value of X. That is, for any value of X that interests us, we can trace a vertical line from that place on the X-axis up or down to the regression line. The level of Y at which our vertical tracing meets the regression line is the predicted Y value for that X value. We can of course make such predictions for all of the cases in our sample, and see how the predicted Y for each case compares to its actual Y value. We call the differences between actual and predicted Y values the errors or *residuals*. In any sample there will be both positive and negative errors, unless r_{xy} is exactly 1 or –1 and all points are precisely on the line. The regression line is chosen to make these errors, as a group, as small as possible; that is the sense in which the line best represents the pattern of points. Because positive and negative errors would cancel each other out when summing the raw errors, the goal of making the errors as small as possible is interpreted in practice as making the *sum of squared errors* as small as possible. This is the *least squares* principle for choosing a regression line.

The line drawn on the scatterplot is informative, but for many purposes we need to instead work with the actual equation representing the line. The line is defined by its *slope* (the change in the predicted value of Y per unit change in X) and its *intercept*, the place at which it crosses the Y-axis, so those two numbers need to be indicated in the equation. We write the equation for the line in terms of predicted Y (or $\hat{Y}$, spoken as "Y hat," where the hat indicates the predicted value). The general form of the equation for the regression line with a single X is then $\hat{Y} = a + b\,X$, and the formulas that statistical software packages use to calculate numerical values of a and b from the sample data are determined by the least squares principle. Once we have numerical values for a and b, we can obtain the predicted value of Y for any X value by plugging that X value into the equation. For example, if our software calculated the regression equation for our sample data as $\hat{Y} = 5{,}320 + 300\,X$, then the predicted Y when X = 10 is 5,320 + (300 × 10), or 5,320 + 3,000 = 8,320.

The slope b is the main object of interest in the regression equation, as it is directly providing information about the relationship between X and Y in our sample. That is, it tells how the predicted Y changes as we change the X value that we are plugging into the equation. Adding one unit to the plugged-in X value will change $\hat{Y}$ by b units: X is multiplied by b in the equation, and $b(X + 1) = bX + b$, so the calculated value of $\hat{Y}$ now differs by b units from what it was before adding 1 to the value of X. For example, if Y represents dollars of income and X represents years of age, b would tell us how the predicted dollars of income change when we plug in a value for age that is 1 year older than we plugged in before. For the regression equation above, in which b = 300, our summary of the information on X and Y in our sample data is that predicted income increases by $300 as the age we plug in increases by 1 year. If b were negative, then the change in $\hat{Y}$ would be a decrease. If b were –400, then the predicted income decreases by $400 as the age we plug in increases by 1 year. Because the regression equation refers to a straight line, this change in $\hat{Y}$ applies whether we are thinking about age 25 vs. age 26, or age 32 vs. 33, and so on. Note that when we refer to a one-unit increase in X, we mean this sort of interpretation of what happens when we plug an X value into the regression equation, not that we are altering our original sample data in some way.

In this way b gives us very specific information on how $\hat{Y}$ relates to X in the sample, and b is often described as the *effect* of X on $\hat{Y}$. The line is steeper when the absolute value of the slope b is larger, and flatter when the absolute value of b is smaller. If b = 0, then the line is perfectly flat, and our prediction of Y would be unaffected by the X value. In the equation, b = 0 would make the (b X) part of the equation zero no matter what X value we plugged in, leaving $\hat{Y}$ = a for any X value. That would occur when there is no relationship between X and Y in the sample. Although this language of "effect" is convenient, we should remember that it does not necessarily imply a true causal effect of X on Y.

In the social sciences, the intercept a in the regression equation is usually of much less interest to researchers than the slope b is. Of course the intercept must be included when calculating predicted values of Y, so it is certainly a necessary part of the equation. But it does not speak to the question of how Y relates to X like b does, and that relationship is usually the researchers' main concern. One specific interpretation of a is as the predicted Y when X = 0: if we plug in X = 0, then (b X) will be zero even if b is not, and we obtain $\hat{Y}$ = a. In some cases it will be interesting to determine the predicted Y when X = 0, but in practice many variables used in social science do not have zero as a legitimate value. For instance, survey data on income will not include data on infants, and when X represents age, X = 0 will not occur in the sample data. Then even if the calculation is straightforward, there is not any meaningful reason to ask what the predicted income is for someone at X = 0. In any event, the direct information on the relationship between X and Y that is provided by b in the regression equation makes it our main focus for interpretation. Note too that in the regression output from most software, the column labeled "b" will include the value of a. a can be distinguished in the output by a term such as "intercept" or "constant" rather than a variable name.

We can calculate the regression equation no matter what the value of r_{xy} may be in our sample, but, as mentioned above, we regard the regression equation as a better representation of the sample data when r_{xy} is large (in absolute value). When r_{xy} is close to zero, the quality of the predictions for cases in the sample generally will be poor, and we will have many large errors or residuals relative to the overall variability in Y. In terms of the plot, there will be many points that have a large vertical distance from the regression line, so the regression line will not be a good fit to the cloud of points. Therefore, the value of r_{xy} is important context that we need to keep in mind when interpreting the specific information provided by the regression equation on the relationship between Y and X.

Note too that for most actual research questions, we will quickly recognize that surely more than just one independent variable would be helpful in predicting Y. In the example of age and income, we immediately think of education as another independent variable that would improve our ability to predict Y. Simple (bivariate) linear regression will therefore seem rather unrealistic in many research settings. Throughout the rest of the book we mainly focus on multiple regression, with more than one X included in the regression equation and hypothesized to have an effect on Y. In most social science applications, the richer multiple regression framework will be preferable to simple regression.

1.3 Example: State-Level Mental Stress in the United States

Throughout this book we have many fully worked-out and discussed examples that apply the methods to data, as well as many exercises meant to reinforce readers' understanding of the analytic techniques. We hope that readers will not only read the examples but also try to reproduce the examples' actual analyses as well as work through the exercises. Reproducing the examples or working on the exercises will require access to some statistical software. As noted in the Preface, any of the software packages that are commonly used in the social sciences R— for example, SPSS, Stata, SAS, or R— can be used for all of the analyses discussed in the book (and many more).

Our first example reviews descriptive statistics, correlation, and simple (or bivariate) regression. It involves a researcher who collected aggregate-level data on mental distress (the percentage of adults who reported that their mental health was not good for 14 or more of the past 30 days), median income (in thousands of dollars), and the percentage of single-parent households from the 50 U.S. states. She also categorized states into five geographical regions (MW, NE, SE, SW, and W). The state-level data collected from the Census Bureau and Centers for Disease Control are as follows:

Data:

State ID	State Name	Percentage Mental Distress	Median Income (in $1,000s)	Percentage Single-Parent Households	Region
1	Alabama	14.4	44.8	11.9	SE
2	Alaska	10.2	74.4	11.6	W
3	Arizona	11.7	51.3	11.6	SW
4	Arkansas	16.4	42.3	11.4	SE
5	California	10.6	63.8	11.5	W
6	Colorado	10.6	62.5	9.5	W
7	Connecticut	10.7	71.8	10.2	NE
8	Delaware	11.1	61.0	11.5	SE
9	Florida	11.4	48.9	10.7	SE
10	Georgia	12.6	51.0	13.0	SE
11	Hawaii	9.2	72.0	9.6	W
12	Idaho	10.7	49.2	9.4	W
13	Illinois	10.0	59.2	10.5	MW
14	Indiana	13.2	50.4	11.2	MW

(*Continued*)

State ID	State Name	Percentage Mental Distress	Median Income (in $1,000s)	Percentage Single-Parent Households	Region
15	Iowa	10.0	54.6	9.4	MW
16	Kansas	9.8	53.6	10.0	MW
17	Kentucky	14.7	44.8	11.3	SE
18	Louisiana	13.1	45.7	14.0	SE
19	Maine	12.6	50.8	9.2	NE
20	Maryland	10.1	76.1	11.6	SE
21	Massachusetts	11.9	71.0	9.7	NE
22	Michigan	13.4	50.8	10.6	MW
23	Minnesota	9.3	63.2	9.3	MW
24	Mississippi	14.1	40.5	15.2	SE
25	Missouri	13.2	49.6	10.6	MW
26	Montana	10.4	48.4	8.5	W
27	Nebraska	9.5	54.4	9.7	MW
28	Nevada	14.2	53.1	12.0	W
29	New Hampshire	12.7	68.5	8.7	NE
30	New Jersey	10.7	73.7	10.3	NE
31	New Mexico	12.5	45.7	12.9	SW
32	New York	10.6	60.7	11.1	NE
33	North Carolina	12.1	48.3	11.5	SE
34	North Dakota	9.0	59.1	8.4	MW
35	Ohio	12.9	50.7	11.2	MW
36	Oklahoma	14.3	48.0	11.6	SW
37	Oregon	13.0	53.3	9.6	W
38	Pennsylvania	12.6	54.9	9.8	NE
39	Rhode Island	13.5	58.4	11.3	NE
40	South Carolina	13.7	46.9	12.1	SE
41	South Dakota	8.3	52.1	10.1	MW
42	Tennessee	13.7	46.6	11.3	SE
43	Texas	10.6	54.7	12.8	SW
44	Utah	11.5	62.5	9.0	W

State ID	State Name	Percentage Mental Distress	Median Income (in $1,000s)	Percentage Single-Parent Households	Region
45	Vermont	11.9	56.1	8.8	NE
46	Virginia	10.9	66.1	10.2	SE
47	Washington	11.4	62.8	9.3	W
48	West Virginia	16.5	42.6	9.6	SE
49	Wisconsin	11.6	54.6	9.7	MW
50	Wyoming	12.1	59.1	9.3	W

After entering the above data in statistical software, the researcher ran the appropriate descriptive analyses for mental distress, median income (in thousands), percentage of single-parent households, and geographical region. Because mental distress, median income, and single-parent households are continuous numerical variables, the mean, median, standard deviation/variance, and range (minimum to maximum values) are all appropriate descriptive statistics. On the other hand, frequencies and percentages are appropriate descriptive statistics for region, because it is a categorical rather than numerical variable. Note that the squared standard deviations in the table do not quite match the corresponding variances, but that is simply due to rounding. Statistical software output will usually include more decimal places when giving these descriptive statistics.

	Mean	Median	Standard Deviation (Variance)	Range (Min-Max)
Percentage mental distress	11.9	11.8	1.8 (3.4)	8.2 (8.3–16.5)
Median income in $1,000s	55.7	54.0	9.2 (84.3)	35.6 (40.5–76.1)
Single-parent households	10.7	10.6	1.4 (2.1)	6.8 (8.4–15.2)

Region	Frequency (f)	Percentage (%)
MW	12	24.0
NE	9	18.0
SE	14	28.0
SW	4	8.0
W	11	22.0
Total	50	100.0

After running the descriptive statistics and becoming familiar with the data set, the researcher decided to explore the possibility that median income in a state influences, or at least predicts, the state's level of mental distress. In this case, median income is the independent variable (X) and mental distress is the dependent variable (Y). Correlation and (simple) linear regression are appropriate tools for examining the relationship between two numerical variables. The results of correlation and simple regression analysis are summarized below, in the output from statistical software:

Correlation Analysis

	Median Income in $1,000s (X)	Percentage Mental Distress (Y)
Median income in $1,000s (X)	1.000	−0.600
Percentage mental distress (Y)	−0.600	1.000

Simple linear Regression Analysis

	b	Standard error of b	t	p-Value
Intercept (constant)	18.625	1.309	14.227	0.000 (<0.001)
Median income in $1,000s (X)	−0.121	0.023	−5.202	0.000 (<0.001)

The correlation between median income (X) and mental distress (Y) is −0.600, indicating a negative and moderately strong linear relationship between those two variables. The simple regression results show a value of 18.625 for the intercept a and −0.121 for the slope b. The regression equation for the relationship between median income (X) and mental distress (Y) in this sample is therefore $\hat{Y} = a + b\,X = 18.625 + (-0.121)\,X$. The b value shows that, from the data in this sample, predicted mental distress ($\hat{Y}$) will decrease by 0.121 units (0.121% of the state's adults) as the value of median income (in $1,000s) increases by one unit (or, said more naturally, the value of median income increases by $1,000). The negative value of b corresponds with the negative correlation, as both indicate a negative relationship between median income and states' levels of mental distress. Note that the software also produced several other statistics that for now we are not interpreting, including the *standard error of b*, t, and the p-value. We will explain those statistics in Chapter 2.

The researcher also believes that the percentage of single-parent households may predict the state's level of mental distress. To explore that, the percentage of single-parent households is set as the independent variable (X), and mental distress remains the dependent variable (Y).

Correlation Analysis

	Percentage Single-Parent Households (X)	Percentage Mental Distress (Y)
Percentage single-parent households (X)	1.000	0.379
Percentage mental distress (Y)	0.379	1.000

Simple Regression Analysis

	b	Standard Error of b	t	p-Value
Intercept (constant)	6.701	1.851	3.621	0.001
Percentage single-parent households (X)	0.488	0.172	2.836	0.007

The correlation between the percentage of single-parent households (X) and mental distress (Y) is 0.379, indicating that there is a positive but somewhat weak linear relationship between those two variables. According to the simple regression results, with intercept a = 6.701 and slope b = 0.488, the regression equation predicting mental distress is $\hat{Y} = a + b X = 6.701 + (0.488) X$. That is, the sample data can be summarized as indicating that a state's predicted level of mental distress ($\hat{Y}$) increases by 0.488 (0.488% of its adults) for each one unit (one percentage point) increase in the percentage of single-parent households.

1.4 Exercise: Suicide Rates in Japanese Prefectures

Japan is known for its high suicide rates compared with other industrialized countries. A government official wanted to investigate possible demographic and environmental factors related to suicide rates. He collected data on the suicide rate per 100,000 population, older adult (65 years and older) population percentage, and the average number of clear weather days per year for 47 Japanese prefectures (administrative units that are roughly like American states), which he also categorized by region. The data collected from the Japan Statistical Yearbook are as follows:

Prefecture ID	Prefecture Name	Region	Suicide	Older Adult Population	Clear Days per Year
1	Aichi	Chubu	19.35	20.13	23
2	Akita	Tohoku	32.97	29.47	9
3	Aomori	Tohoku	29.35	25.71	7
4	Chiba	Kanto	21.38	21.24	30

(*Continued*)

Prefecture ID	Prefecture Name	Region	Suicide	Older Adult Population	Clear Days per Year
5	Ehime	Shikoku	20.89	26.48	26
6	Fukui	Chubu	19.85	24.94	21
7	Fukuoka	Kyushu/ Okinawa	23.13	22.14	27
8	Fukushima	Tohoku	25.04	24.84	14
9	Gifu	Chubu	20.47	23.98	37
10	Gunma	Kanto	25.30	23.46	34
11	Hiroshima	Chugoku	21.22	23.66	[illegible]
12	Hokkaido	Hokkaido	25.30	24.66	8
13	Hyogo	Kansai	22.48	22.92	17
14	Ibaragi	Kanto	23.60	22.39	40
15	Ishikawa	Chubu	22.31	[illegible]50	13
16	Iwate	Tohoku	32.03	[illegible].07	8
17	Kagawa	Shikoku	21.59	25.40	25
18	Kagoshima	Kyushu/ Okinawa	24.2[illegible]	26.38	32
19	Kanagawa	Kanto	[illegible].8[illegible]	20.11	26
20	Kochi	Shikoku	25.79	28.53	34
21	Kumamoto	Kyush[illegible]/ Okina[illegible]	24.82	25.48	31
22	Kyoto	[illegible]ansai	22.15	22.99	18
23	Mie	Kansai	18.92	24.10	37
24	Miyagi	Tohoku	22.70	22.19	9
25	Miya[illegible]aki	Kyushu/ Okinawa	27.05	25.64	49
26	N[illegible]gano	Chubu	23.28	26.44	18
27	Nagasaki	Kyushu/ Okinawa	25.79	25.86	41
28	Nara	Kansai	19.06	23.84	25
29	Niigata	Chubu	28.43	26.16	9
30	Oita	Kyushu/ Okinawa	22.22	26.48	26
31	Okayama	Chugoku	20.67	24.94	29
32	Okinawa	Kyushu/ Okinawa	25.34	17.30	15

Prefecture ID	Prefecture Name	Region	Suicide	Older Adult Population	Clear Days per Year
33	Osaka	Kansai	23.64	22.14	18
34	Saga	Kyushu/ Okinawa	25.88	24.47	39
35	Saitama	Kanto	22.82	20.36	51
36	Shiga	Kansai	21.83	20.48	26
37	Shimane	Chugoku	25.66	28.87	17
38	Shizuoka	Chubu	22.68	23.69	42
39	Tochigi	Kanto	24.65	21.81	29
40	Tokushima	Shikoku	19.36	26.75	23
41	Tokyo	Kanto	21.48	20.08	31
42	Tottori	Chugoku	24.62	26.15	15
43	Toyama	Chubu	22.78	26.08	18
44	Wakayama	Kansai	24.85	27.05	25
45	Yamagata	Tohoku	26.26	27.54	8
46	Yamaguchi	Chugoku	24.05	27.91	28
47	Yamanashi	Chubu	27.00	24.57	33

1. After entering the above data in statistical software, use the software to calculate appropriate descriptive statistics for region, suicide rate, older adult population, and the number of clear days per year.
2. Use the software to perform correlation and simple linear regression analyses investigating the relationship between the older adult population and the suicide rate. Interpret the results as fully as possible.
3. Use the software to perform correlation and simple linear regression analyses investigating the relationship between the number of clear days per year and the suicide rate. Interpret the results as fully as possible.