Regression analysis is, by far, the most commonly employed statistical modeling strategy in the social and behavioral sciences. At the same time, it is also an important tool for forecasting and decision-making in a variety of professional fields. This book will provide an introduction to the theory, methods, and practice of regression analysis. Our goal is to provide readers with the skills that are necessary to: (1) Understand, and evaluate the professional literature that uses regression analysis; (2) design and carry out studies that employ regression techniques for testing substantive theories; and (3) prepare to learn about other statistical procedures that build upon a foundation provided by the general linear model.

What is regression analysis? Regression analysis can be defined in many different ways. Indeed, we will provide a number of definitions throughout the course of this text. But, for now, we can use one that is brief and simple: Regression analysis is a tool for describing relationships among variables. This brief definition is accurate. But, it definitely bypasses an enormous amount of detail. We believe it is helpful for readers to have some sense of the “big picture” before we jump directly into these details. Therefore, we will start with a very simple example that illustrates the general idea of regression analysis and introduces many of the basic topics that we will address throughout the rest of the book.

Table 1 shows data for thirty observations on two variables. For present purposes, the substantive content of the data is unimportant, so we will merely label the variables “X” and “Y”. What can we say about any structure that might exist within these data? By the standards of real-life scientific research, this is certainly not a large dataset. Nevertheless, most people would find it difficult to comprehend the information in its “raw” form— that is, as it is shown in Table 1.

As an alternative approach, we can construct a graph to convey the same information. In order to do so, we will draw two coordinate axes which cover the range of values on X and Y. In other words, we will place two number lines perpendicular to each other as shown in Figure 1a. Now we begin locating points to represent the observations relative to these
axes. The point locations are determined by the values of the variables for each observation. Consider the first observation, which has a value of 5 on $X$ and 8 on $Y$. The dashed line segments in Figure 1b show how we can project upward from the 5 point on the horizontal axis (which, you will recall, represents the values of the variable, $X$) and sideways from the 8 point on the vertical axis (which represents the variable, $Y$). At the intersection of these two projections, we draw a point. And, having done so, we can remove the dashed line segments (as shown in Figure 1c). Now, this point represents the information we possess about observation 1, since its location is completely determined by the data values for that observation.

We repeat this process—projecting from the axes at the respective variable values and inserting a point at the intersection of the projections—for the remaining 14 observations from Table 1. The result is shown in Figure 1d. As you may already know, this type of graph is called a *scatterplot*. And, generalizing from what we said earlier about the first observation, the scatterplot depicts all of the information that was contained back in Table 1. Note that we cannot tell immediately from the graph which points correspond to which rows in Table 1. But, this is no real problem because we are more interested in the overall structure of the entire dataset, rather than the data values associated with individual observations.

The advantage of the pictorial representation in Figure 1d is that we can see immediately a feature in the scatterplot that was not particularly apparent in the numerical version of the dataset. Stated informally, the values of the two variables “go up and down together.” More precisely, observations that have small values on $X$ also have small values on $Y$, observations with large $X$ values also have large $Y$ values, and so on. This nonrandom correspondence in the relatively sizes of the data values across $X$ and $Y$ is manifested in the graph by the “point cloud” which extends from lower left to upper right within the plotting region.

The first task of regression analysis is to provide an objective standard for determining when systematic structure, such as that shown in Figure 1d, actually exists within a dataset. What would the data look like if no such structure exists? Figure 2 provides an example. It
shows a scatterplot created from the dataset given in Table 2. Here, there is no discernible pattern in the way that values of \( X \) and \( Y \) correspond within observations. In other words, observations with small \( X \) values exhibit \( Y \) values that range across the entire range of the \( Y \) variable. And, the same is true for observations with medium, and large \( X \) values. The result is a “shapeless” point that extends fairly evenly throughout the entire plotting region in Figure 2.

The differences between Figure 1d and Figure 2 are quite stark. So, one could reasonably ask why we need a special data analysis tool for distinguishing between them. One answer is that, for various reasons, the structures that exist within social scientific data are often relatively weak. Therefore, it is often difficult to perceive those structures through simple visual inspection of a scatterplot. Given the subjective nature of human perception, it is possible that an analyst could either overlook the existence of a weak, but still nonrandom, pattern or “see” a structure that is really nothing more than a random configuration of points. Regression analysis provides a means of bypassing this subjectivity with a set of relatively objective standards for assessing the existence of structure in bivariate (or eventually multivariate) data.

The second task of regression analysis is to provide a quantitative description of any nonrandom structure that exists within the data under consideration. For example, Figure 3 presents a larger version of Figure 1d. Here, we can discern that the rate of increase across the values of the variables is relatively constant. To see what we mean by this, scan across the \( X \) axis of the scatterplot from left to right. As we move a fixed distance in the horizontal direction, the average movement upward in the point locations along the vertical axis remains constant. So, consider the interval from 5 to 10 along the \( X \) axis. At the left side of this interval (i.e., around the value of 5), the average \( Y \) value for the points is 8. At the right side of the interval (i.e., around the value of 10), the average \( Y \) value for the points is 12. Thus, we could say that a difference of 5 units along the horizontal axis corresponds to an average difference of 4 units (in the positive direction) along the vertical axis. Now
consider another identically-sized interval along the X axis—say, from 20 to 25. Here, the average Y values increase from 20 at the left side of the horizontal interval to 24 at the right side. Once again, this is an average difference of 4 units along the Y axis. And, it turns out that any five-unit interval along the horizontal axis will correspond to an average difference of 4 units along the vertical axis. This consistent increase in the average Y values, as we move from left to right across the scatterplot, implies that the overall point cloud has a shape that is straight rather than curved.

Extending this reasoning, we might try to superimpose a curve on the point cloud in such a way that this curve approximates the general shape of the point cloud itself. Here, the shape of the cloud in Figure 3 suggests that the most approximate "curve" is a straight line. Speaking informally, we want to pass such a line through the middle of the point cloud such that it comes as close as possible to as many of the points as possible. Figure 4 shows just such a line superimposed over the points. Notice how this line captures the predominant structure that we noticed in our earlier visual inspection of the scatterplot. The deviations of the various points from the line are all relatively trivial, compared to the major pattern of the point cloud, which is captured by the line itself. Thus, we might say that the line actually provides a summary description of the data. Regression analysis is the tool that we employ for positioning this line relative to the plotted points.

Going even further, the linear structure within the data can, itself, be represented very succinctly with the following equation:

\[ Y_i = 4 + 0.8X_i + e_i \]

Do not worry too much about the meaning of this equation at the present time. Eventually, when we construct similar equations for other, more meaningful, datasets we will try to impose substantive interpretations on their components. For now, we will merely point out that regression analysis is a method for finding the specific equation to be associated with a particular dataset. In fact, regression seeks the best equation for a given dataset in the sense of finding the curve (again, a straight line in this case) that provides the closest
possible approximation to the plotted points.

The third task of regression analysis is to assess how well the fitted curve (again, a line in this case) actually describes the data. Figure 5 illustrates what we are talking about here. The panels of the figure show three scatterplots. For all three cases, the same line provides the “best” description of the plotted data. But, you can easily see that the quality of this description improves as we move from panel A to panel C. In the first graph, points are widely scattered throughout the plotting region; many of them lie in positions that are quite distant from the fitted line. In the second panel, the point cloud is still spread fairly widely across the graph, but not as much as in the first case. Finally, the data points in the third panel all lie very close to the line. We would like to obtain a single numerical value that summarizes this general quality; that is, the degree of consistency between the fitted line and the data points. Such a summary value is usually called a “goodness of fit” statistic.

Fourth, we will frequently want to use regression analysis to analyze relationships among more than two variables. For example, consider Table 2 which adds a new variable, Z, to the data that were originally presented in Table 1. Does variable Y show any systematic pattern with Z? Any answer to this question is complicated by the fact that we already know that Y is related to X. For reasons that we will eventually explain, it is generally not appropriate to simply examine the separate relationships between Y and X and between Y and Z, respectively. Therefore, it is more important to ask whether Y shows a pattern with Z even after we have taken into account the nonrandom structure involving X. And, of course, we will want to know whether the previously observed relationship between X and Y continues to exist after we take Z into account.

One strategy for answering these concerns would be to generalize the approach that we have been using so far. To this end, Figure 6 shows a three-dimensional scatterplot. It is created by adding a third perpendicular axis, representing Z, to the two-dimensional scatterplot that was depicted in Figures 1D and 3. Points are located within the three-dimensional space by projecting from the axes just as we did earlier. We are now using data values from
three variables, rather than two, to define the projections. Of course, the page itself is a two-dimensional flat surface. So, we must use the artist’s trick of perspective drawing in order to simulate the resultant three-dimensional coordinate space. Once again, we can see that there is a nonrandom pattern in the placement of the points. Generally speaking, the point cloud is shaped like a flat “slab” that is tilted within the three-dimensional plotting region. The general orientation of this slab indicates that observations with small values on $Y$ (i.e., those that project to relatively low positions on the vertical axis) also tend to have small values on $X$ and $Z$ (i.e., they occur near the front corner of the plotting region). Conversely, observations with large $Y$ values (i.e., high positions along the vertical axis) tend to have large values on $X$, $Z$, or both (i.e., they are located closer to the far corners of the plotting region). As a way of summarizing the structure among these three variables, we can insert a flat surface that passes through the middle of the point cloud. This is shown in Figure 7. Just as in the two-variable (or bivariate) case, this surface is oriented to come as close as possible to as many points as possible. Regression analysis provides tools to find the proper orientation for this surface. It also produces an equation like the following to describe the fitted surface, as well as the point locations relative to that surface:

$$Y_i = 4 + 0.8X_i + 0.5Z_i + e_i$$

Once again, we will defer interpretation of this equation until later in the book. For now, simply note that it explicitly relates the values of the $Y$ variable for any observation (denoted as $i$ in the equation) to the values of $X$ and $Z$ for that same observation. As we will see, the term designated $e$ on the righthand side of the equation stands for the distance between the surface and the plotted points for observation $i$. What if we want to examine the relationship between $Y$ and other variables along with $X$ and $Z$? At this point, the visual representation becomes problematic because it is very difficult (to say the least!) to draw scatterplots in four or more dimensions. Conceptually, however, there is no problem in thinking about higher dimensioned geometry. And, perhaps more to the point, we can incorporate additional variables into the analysis by simply including more terms on the
righthand side of the equation. Situations in which a single $Y$ variable is related to more than one other variable are called multiple regression analysis.

The fifth task of regression analysis arises from the fact that we will typically want to generalize our conclusions beyond the data that are immediately observable. To use common statistical terminology, we only observe information in a sample. We assume that the observations within this sample are drawn from some broader population. And, it is the characteristics of this population that are really of interest to us. Therefore, the sample data are not of intrinsic interest, themselves; instead, they are only a means to a broader end.

The problem is that sample data always represent incomplete information. This, in turn, creates an element of uncertainty about any general conclusions that we might try to draw from our analysis. The conundrum is that we would like to construct a scatterplot of the two variables using all elements of the population. However, this is always impossible because populations are too large and too widely-dispersed in space and/or time. If nothing else, practical considerations prevent us from observing the entire population. Instead, all we have is a sample representing a fraction (and usually a very tiny fraction) of the overall population. Of course, we hope this sample is a microcosm of the population. In other words, we want the characteristics of the sample to mirror those of the population. But, even if we achieve this general objective it is likely that the correspondence between sample and population will not be exact. Furthermore, there is always the possibility that our sample is a relatively poor representation of the population. Through the accidental selection of nonrepresentative observations, we could end up with a sample whose characteristics are not very consistent with those of the population from which it is drawn. In such cases, any conclusions we draw from our observations could be misleading or even wrong. And, since we cannot observe the population, we will never know for sure whether this is the case.

At this stage, it might sound like we are stuck between a rock and a hard place. Fortunately, however, statistical theory comes to our rescue. It enables us to quantify the uncertainty that exists in our conclusions about the underlying structure of the population.
which are based only upon our observation of the sample. In a sense, we will be adding caveats and qualifications to our interpretations, due to the fact that we are always stuck with inherently incomplete information.

The sixth task of regression analysis is to diagnose and try to “fix” any problems that might arise over the course of carrying out the preceding five tasks. For example, we have assumed so far that our data conform to linear patterns. But, what if the relationship between two variables exhibits the kind of structure shown in Figure 8? (show a curved pattern) From panel A it is clear that the point cloud is “curved.” Any attempt to provide an accurate description of the data must take this into account. If we insist on superimposing a line over the data, as in Figure 8B, it is easy to see that it just does not follow the contour of the point cloud. Hence, it would misrepresent the structure of the data. However, when this situation arises, regression techniques can be used in a straightforward manner to fit an appropriate curve to the data, as shown in Figure 8C. The latter does follow the predominant trend from the plotted points. So, it does provide a reasonable description of the structure within the data.

In a slightly different vein, the whole process of statistical inference (i.e., generalizing to populations from samples) only works if we are willing to make a series of fairly specific assumptions about the processes that generated the observed data in the first place. To the extent that these assumptions are violated, then the validity of any conclusions we draw from our analyses will be compromised. Therefore, a critical element of regression analysis involves careful consideration of these assumptions and whether they are actually met in the data at hand. If we decide that any of these assumptions are violated, then we can usually take corrective actions to overcome the problems.

From the preceding discussion, you might get the impression that regression analysis is carried out as a sequential series of discrete steps. However, let us close this section by stating emphatically that this is not the case. The ordering of the tasks is not written in stone. The important idea is that regression analysis confronts a variety of concerns more or
less simultaneously. Each study poses its own problems and challenges. The great strength of modern regression analysis is that it can be adapted to deal with an enormous array of data analytic situations. And, that contributes directly to its utility as a research tool.

**SUBSTANTIVE THEORY AND STATISTICAL MODELS**

The discussion in the preceding section provided a brief explanation of what regression analysis can do. Specifically, it is a tool for evaluating relationships among variables. But, why would we want to perform a regression analysis and examine such relationships in the first place? We believe there are two major answers to that question: (1) Theory construction and (2) forecasting. Let us consider each of these in turn.

**Theory Testing**

As scientists, we try to create theories which account for interesting phenomena drawn from our respective substantive areas of expertise. For present purposes, we can define a theory as a statement that describes how variables coincide or “go with” each other in the world around us. Essentially, theories are propositions about cause and effect relationships. In this capacity, they try to explain how things work. Of course, the content of theories (and, what we mean by “things”) will vary by substantive field. Some simple examples of theory-based statements include the following:

- Domestic economic distress within nations leads to aggressive behavior toward other nations.
- Good nutrition facilitates public health.
- Social isolation heightens individual alienation.
- Individual partisan affiliations affect attitudes on particular political issues.
- Population concentrations within urban areas are affected by the availability of public transportation facilities.
- Administrative professionalism affects the efficiency of operations within bureaucratic organizations.
Each of the preceding statements comprises a simple theory. And, in every case, we can look beyond the specific substantive content to observe that the theory makes an assertion that one variable (domestic economic distress, good nutrition, social isolation, etc.) leads to or “causes” another variable (aggressive behavior toward other nations, public health, alienation, etc.). The process of scientific research involves holding these causal assertions up against empirical evidence in order to determine whether the theory itself provides an accurate description of reality.

The nature of causality is a difficult issue. In fact, philosophers of science continue to disagree about what it means to say that one thing causes another. But, there does seem to be a consensus about the empirical manifestations of causality. Specifically, if variable $X$ causes variable $Y$, then the following conditions must hold:

1. **Precedence**: $X$ must occur before $Y$. It makes no sense to have an effect which precedes its own cause.

2. **Covariation**: Variability in $X$ must coincide with variability in $Y$. In other words, differences in the amount of $X$ that are present must correspond in some systematic, identifiable way with differences in the amount of $Y$ that is present.

3. **Isolation**: The observed covariation between $X$ and $Y$ must be nonspurious. That is, the structural connection between the two variables must continue to exist even after the potential effects of other variables are explicitly taken into account.

An effective research design would enable the analyst to determine whether these three conditions exist with respect to the $X$ and $Y$ variables that are actually under investigation. One approach would be to use an experiment. In this case, the researcher would randomly sort the available observations into two or more groups. Then, within one of the groups, the amount of $X$ would be manipulated systematically. Finally, the level of $Y$ is measured within each of the groups. If $X$ truly causes $Y$, then the manipulation should have a discernible impact on the amount of $Y$ that is observed within the manipulated group— that is, it should differ systematically from the amount of $Y$ that is observed within the non-manipulated group. In an experimental study, the precedence condition is handled by the researcher’s overt manipulation of the hypothesized causal variable. The isolation condition
is handled through the random assignment of observations across the groups; if carried out properly, this holds constant all possible causes of $Y$ apart from $X$. Finally, covariation is typically measured with a procedure called Analysis of Variance (ANOVA) which assesses the difference in the average level of $Y$ across the groups.

Experimental designs constitute an excellent strategy for making causal inferences. But, there are many research contexts where experiments are either impractical or just impossible. In those cases (which arise very frequently in the social sciences), regression analysis provides a viable alternative. It can be used to test assertions about causal relationships using observational data that are drawn directly from the external world. First, regression analysis provides an empirical measure of covariation between variables $X$ and $Y$. In fact, this measure is based upon a specific statistical summary of the data called the covariance.

Second, regression analysis handles the isolation condition, not by holding other variables constant, but rather by statistically removing the effects of other variables from the measured covariance between $X$ and $Y$. Note, however, that the analyst must specify exactly which other variables must be taken into account during the course of the regression analysis. This is different from an experiment where the design implicitly controls for the effects of all other influences on $Y$.

Finally, regression analysis does not really provide a way of dealing with the precedence condition. Instead, it is up to the analyst who must use substantive insights and prior knowledge to make sure that $X$ truly does precede $Y$.

One might conclude that regression is a weaker tool for assessing causality than an experimental design. After all, with regression we assert precedence. In an experiment, the researcher can guarantee it. In regression we have to specify the control variables necessary to isolate the effect of $X$ and $Y$, while that is handled automatically in an experimental framework. But, it is important to emphasize that experiments have their own weaknesses, such as problems in the group assignment, the manipulation of the independent variable, and the generalizability of any statistical results. Therefore, we reject the notion that ex-
Experimental designs analyzed with ANOVA are any better or worse than studies based upon observational data analyzed using regression methods. Of course, the researcher must always be aware of the strengths and weaknesses inherent in each approach. Beyond that, they merely represent different strategies to testing causal inferences.

Think back to the set of example theories listed earlier. Most of them would be difficult to test using purely experimental methods. Social scientists simply cannot manipulate phenomena like the economies of nations, the partisan affiliations of individual citizens, or the professional backgrounds of administrators. One could also raise ethical questions about manipulating factors like nutritional intake of population groups or the degree of interpersonal contact among experimental subjects. In such cases, regression analysis is really the only viable strategy for testing theoretical assertions. And, we believe that this kind of situation arises more frequently in the social sciences than do research questions that are amenable to experimentation. This alone guarantees that regression analysis is an important tool for social scientific research.

**Forecasting**

Researchers in non-academic, applied settings (e.g., government agencies, businesses, non-profit organizations, etc.) are usually less concerned with theory testing than they are with predicting the future. Some examples include the following:

- A state director of human services needs to know how much money to allocate for welfare programs in upcoming budget cycles.
- A private company’s marketing manager wants to determine how a product should be distributed across its retail outlets in order to generate optimal consumer response.
- A nonprofit organization serving a population with special needs would benefit from information about the likely growth and/or constriction of its constituency.
- Local school board officials require information about the future enrollments so they can allocate resources to meet students’ needs.
- Federal health officials need to make projections about the spread of infectious diseases to make sure that adequate medical supplies are available.

12
In each of the preceding cases, the researcher wants to predict the level of a relevant variable (i.e., monetary allocations for welfare, supplies of company products, constituency sizes, etc.) in order to achieve an immediate objective (i.e., financing welfare programs, optimizing consumer responses, meeting population needs, and so on). In this case, the focus is on a tangible outcome. With theory testing, interests center more on the nature of the underlying structure that generates those outcomes.

Confronted with a prediction problem of this sort, how might a decision-maker proceed? In most cases, they would probably consider previous occurrences of the phenomena they are investigating. As a concrete example, the state director of human services might look at welfare expenditures during earlier fiscal years. Given the distribution of spending values, an intuitive approach would be to use the mean value as the “single best guess” for a future expenditure. That would certainly be an efficient way to employ information about previous state welfare expenditures.

But, the director probably has even more information to bring to bear on the problem. He/she may notice that expenditures in a given year are related to the size of a state’s low-income population during the preceding year. In other words, years during which the number of poor people within the state is relatively low are followed immediately by years with relatively small welfare expenditures, and vice versa. Accordingly, the director would probably begin by looking at the size of the current low-income population. He/she could then look at the average welfare expenditures during previous years when the low-income population was a similar size.

Notice that in the later case, the director is using a conditional mean rather than an overall or marginal mean of the relevant variable. The justification for doing so is that the latter would be likely to produce a smaller prediction error than the former. Stated differently, the prediction itself is probably going to be more accurate. If the director can incorporate other variables that seem to be related to welfare expenditures (e.g., the state’s economic conditions, changes in welfare eligibility, etc.) then it may well be possible to
reduce the prediction error even further.

Regression analysis is a tool that is useful for precisely this kind of situation. If a relevant variable (again, $Y$) is related to one or more other variables (say, $X_1$, $X_2$, and $X_3$), then regression analysis enables us to make an optimal prediction about the value of $Y$, given specific combinations of values for $X_1$, $X_2$, and $X_3$. Of course, this prediction will probably not be perfect. In other words, there will be some discrepancy between what we forecast and what actually occurs. But, regression analysis estimates the likely size of such prediction error, at least assuming that the future structure of the relationships between $Y$ and $X_1$, $X_2$, and $X_3$ is identical to that which existed in the past. In this manner, regression analysis not only provides the forecast, itself. It also assesses the magnitude of the uncertainty surrounding that forecast.

At this point, we need to issue a slight disclaimer. From the discussion so far, you might get the impression that theory testing and forecasting are qualitatively different things. However, it is important to emphasize that the two are intimately related to each other. On the one hand, theories are only vindicated if they can predict phenomena that have not yet been observed. On the other hand, the most efficient forecasts occur when the analyst has a theory about the causes of the phenomena under investigation. For these reasons, the distinctions between theory testing and forecasting are differences of emphasis, rather than differences of kind. And, to be honest, most research—regardless whether it occurs in an academic or applied setting— involves elements of both.

**EMPIRICAL RELATIONSHIPS BETWEEN VARIABLES**

Regardless of the ultimate objective, the researcher wants to know why a specific variable varies. To introduce an example that we will use throughout the remainder of the book: Why do some of the American states exhibit large per capita expenditures on health care programs while others devote fewer resources to this policy area? The researcher answers this question by proposing that this variable is related to, or dependent upon, at least one other variable. In generic terms, the former is usually called the dependent variable, and it is denoted as $Y$. 


The later is usually called the independent variable, and it commonly denoted as $X$. If there is more than one independent variable, the names are usually shown as $X$ with a subscript; hence the general nomenclature for three independent variables would be $X_1$, $X_2$, and $X_3$.

Getting back to our ongoing example, we could hypothesize that state health care expenditures are related to the size of the population that is at risk for medical problems. We might be interested in this because we are constructing a theory of governmental policymaking, or we may want to predict state budgetary allocations. In either case, regression analysis provides an approach for measuring the relationship between population size and program expenditures— or between the $X$ and $Y$ variables that are appropriate for any other research context.

Measuring the empirical relationship between variables is important for theory testing: If no such relationship exists, then one of the underlying conditions for causality is violated. And, the theory, itself, must be revised or discarded. The existence of a relationship is also important for forecasting purposes. If the variables under investigation are related, then it should be possible to make predictions with minimal error. If the variables are not related, then the size of the prediction errors would probably be very large, and this would signal that we need to seek another variable— a different $X$ that is more strongly related to $Y$— in order to improve our forecast. In both of these contexts, the “relatedness” of $X$ and $Y$ is critical to the task. For this reason, we need to consider in greater detail what we mean when we say that two variables are related to each other.

**The Meaning of a “Relationship” Between Variables**

How would you define the existence of a relationship between two variables, $X$ and $Y$? Many people would say that $X$ and $Y$ are related if the observations on these variables form a roughly linear pattern when graphed in a scatterplot. Such an answer is not really incorrect: It is true that related variables should exhibit some nonrandom structure when their values are plotted against each other. However, that answer does not go far enough for our purposes. First, the contents of the scatterplot are a consequence of the relationship,
rather than the relationship itself. We want to define “relationship” in such a way that it is independent of the display mechanism used to depict it. Second, a linear pattern is only one type of relationship. While such linear relationships will occur a great deal of our attention throughout the rest of this book, we still want to allow for a broader variety of possible connections between $X$ and $Y$. Third, the preceding definition defines a relationship in symmetric terms. In other words, the statement about the linear pattern in the scatterplot does not distinguish between the independent and dependent variables. This is problematic because our interest is in explaining variance within the dependent variable. For the most part, we will regard variability among the independent variables as a given. It is not something that we seek to explain with a theory or predict through a forecasting model. Fourth, we would prefer a definition that can be extended to allow for several $X$ variables, all of which may be related to $Y$. It is very difficult to incorporate more than two variables in a traditional, two-dimensional scatterplot. For all of these reasons, we will take a different approach in order to define the existence of an empirical relationship.

Before getting to our definition, let us introduce some important ideas. First, the probability distribution of the random variable, $Y$, can be defined as the enumeration of all possible $Y$ values along with the probability that each value occurs. Of course, probabilities are theoretical quantities which are not directly observable in the empirical world. Therefore, it is convenient to define the analogous concept of the frequency distribution for $Y$ as the enumeration of $Y$ values along with the empirical frequency of occurrence for each one. The frequencies could be expressed in relative terms, that is as proportions of the total number of observations. In that case, the relative frequency distribution might be interpreted as an empirical estimate of the unobservable probability distribution. If we are thinking about a population, then we will assume it contains $N$ objects, each of which has a value on $Y$. These can be denoted as $Y_1, Y_2, \ldots Y_i, \ldots Y_N$. If we are dealing with sample data, we will assume that there are $n$ distinct observations on $Y$, shown as $y_1, y_2, \ldots y_i, \ldots y_n$. Throughout the book, both the context and the notation should make clear whether we are referring to
a sample or to the population from which the sample is (or could be) drawn. Generally, a distribution can be described using four characteristics:

- The central tendency of the variable’s values.
- The dispersion of the variable’s values.
- The shape of the distribution.
- Outliers.

The central tendency of a distribution can be regarded as the most typical value of the variable; it is the single number that “best” exemplifies the full set of values that occur within the population or the data at hand. It is usually measured by the mean. Dispersion represents the amount of “spread” in the variable's values. It is often measured by the variance which represents the average size of the squared difference between individual data values and the mean value for the distribution. As we will see the variance has nice mathematical and statistical properties. However, it is sometimes difficult to interpret because it is expressed in squared units of measurement. For that reason, the square root of the variance or the standard deviation, is often used to describe the dispersion of a distribution. For certain theoretical distributions (i.e., the normal, t, F, and so on), the shape of the distribution is defined in terms of an equation called the density function. But, with empirical distributions, shape is often assessed a bit more informally by considering two major characteristics: The number of modes (i.e., the modality) and the relative sizes of the upper and lower tails (i.e., symmetry or asymmetry) in the data. Finally, outliers are unusual observations. By definition, there can be no more than a few in any distribution. And many distributions contain no outliers at all. When they exist, outliers are identified individually as observations with unusually extreme data values which are widely-separated from the rest of the observations.

Once again, our attention is focused on the variance of $Y$. Implicitly, we want to determine why the values of $Y$ are dispersed above and below the mean. While not absolutely necessary, this task is a bit more straightforward when the distribution of $Y$ is well behaved, that is $Y$’s distribution is unimodal and symmetric. In that case, the location of the distribution’s
center as well as the interpretation of the variance, are both clear-cut and unambiguous. Similarly, we prefer distributions without outliers because the latter can sometimes cause misleading values for calculating summary statistics on the distribution. For the moment, however, we will not worry about any of these potential problems. But, we will encounter all of them again later!

We next assume that each object or observation with a value on $Y$ also possesses a value on the independent variable, $X$. So, in the population, the $X$ values could be shown as $X_1, X_2, \ldots X_i, \ldots X_N$. In the sample, they would be $x_1, x_2, \ldots x_i, \ldots x_n$. For the moment, let us focus on the population. Assume that variable $X$ takes on $m$ distinct values. We can, therefore, sort the population objects into $m$ distinct groups based upon their $X$ values.

Now, let us return to the $Y$ variable. The marginal distribution of $Y$ is the probability distribution based upon all $N$ objects in the population. This can be divided into $m$ conditional distributions for $Y$ created separately within each of the subgroups defined by distinct $X$ values. For example, $\mu_Y$ is the mean of the marginal $Y$ distribution and $\mu_Y|X_i$ is the mean of the conditional $Y$ distribution calculated only across those observations for which $X = X_i$. There will be $m$ distinct $\mu_Y|X_i$’s, since there are $m$ different subgroups corresponding to the distinct values on the variable, $X$. Of course, each conditional distribution will also have its own conditional variance, shown as $\sigma_Y^2|X_i$, its own shape (i.e., some might be symmetric, while others are skewed, and so on), and maybe even its own outliers (i.e., values of $Y$ that are relatively unusual compared to most of the $Y$ values that coincide with $X_i$).

At last, we are ready to state our definition for the existence of a relationship between variables. We say that $Y$ is related to, or functionally dependent upon, $X$ if the conditional distributions of $Y$ differ in some systematic way across the distinct values of $X$. The precise nature of these differences depends upon the context; they could involve any aspects of the conditional distributions. But, it is safe to say that we will usually focus on differences in the conditional means of $Y$. And, to keep things as simple as possible, we can (at least for
the moment) assume that the shapes of the respective conditional distributions are constant and that there are no outliers within any of the conditional distributions. In that case, we could re-express our definition by stating that $Y$ is related to $X$ if two conditions hold simultaneously: (1) $\mu_{Y|X_i}$ differs from $\mu_Y$ for at least some of the $X_i$’s; and (2) the various $\mu_{Y|X_i}$’s differ among themselves in some discernible, systematic way.

Note how our definition subsumes the initial, incomplete definition of an empirical relationship. The earlier definition is inherently implied by our definition because systematic variability in the conditional $Y$ distributions will be manifested as a nonrandom shape in the cloud of points formed by graphing $Y$ against $X$ in a scatterplot. At the same time, our definition addresses all four of the limitations that we described earlier. First, the relationship is defined as the functional dependence of the $Y$ on $X$; it is not tied to any particular strategy for displaying or conveying this dependence. The variability in the conditional $Y$ distribution can be represented numerically, through the differences across the conditional means, $\mu_{Y|X_i}$, for $X_1, \ldots, X_N$. Alternatively, it can be represented graphically, using a scatterplot in the traditional manner. Again, our definition focuses on the actual dependence of $Y$ and $X$ not the way it is represented.

Second, our definition of functional dependence easily incorporates relationships that are nonlinear in form. A linear relationship implies that the differences between successive conditional $Y$ means are all equal across the entire range of $X$ values. Using our notation for the variables, this can be shown as follows:

$$\mu_{Y|X_1} - \mu_{Y|X_2} = \mu_{Y|X_2} - \mu_{Y|X_3} = \ldots = \mu_{Y|X_{m-1}} - \mu_{Y|X_m}$$

If we were to graph the conditional $Y$ means against the $X$ values, the linear nature of this relationship would be obvious immediately because the array of points would fall along a straight line. But, nonlinearities exist whenever the preceding equalities do not hold and the differences in the successive conditional $Y$ means vary in some systematic way. For example,
the sizes of these differences could increase along with the sizes of the $X$ values? That is:

$$\mu_{Y|X_1} - \mu_{Y|X_2} < \mu_{Y|X_2} - \mu_{Y|X_3} < \ldots < \mu_{Y|X_{m-1}} - \mu_{Y|X_m}$$

This is an example of a nonlinear relationship. Again, the easiest way to see this would be to graph the conditional $Y$ means against the $X$ values. In that case, the array of the plotted points would curve upward as we move from left to right across the horizontal axis of the graph. Of course, if the conditional means vary this implies that the conditional distributions are varying as well. Therefore, this situation fits completely within our definition of a relationship even though it is nonlinear in form.

Third, our definition clearly places greater emphasis on $Y$ than it does on $X$. Again, we are focusing on differences in the conditional $Y$ distribution. This pushes our attention directly toward the phenomenon of interest—the variability in our dependent variable. In contrast, so long as $X$ does in fact vary, the nature and properties of its distribution are largely irrelevant for assessing the relationship between the variables.

Fourth, our definition can be generalized very easily to incorporate multiple independent variables. We simply make the $Y$ distributions conditional upon the values of several independent variables, simultaneously. Assume that we have a total of $k$ independent variables. Our definition would now state that the conditional distributions of $Y$ vary systematically across distinct combinations of values for variables $X_1, X_2, \ldots, X_k$. Once again, we can focus on the conditional means, rather than the entire conditional distributions. The value of an arbitrary observation (say, observation $i$, where $i$ can range from 1 to $n$) on an arbitrary independent variable (say, $X_j$, where $j$ can range from 1 to $k$) is shown as $X_{ji}$. And, the conditional $Y$ mean would be given as $\mu_{Y|X_{1i}, \ldots, X_{ki}}$. Now, $Y$ is functionally dependent upon (i.e., related to) the set of independent variables, $X_1, X_2, \ldots, X_k$ when two conditions hold:

1. The marginal $Y$ mean does not equal the conditional mean for at least some subset of combinations of $X_j$ values, or $\mu_Y \neq \mu_{Y|X_{1i}, \ldots, X_{ki}}$ for at least some $i$. 

20
2. The conditional means of $Y$ vary in some systematic way across the possible combinations of $X_{ji}$

Of course, when there is only a single independent variable (i.e., when $k = 1$), the preceding definition reverts to our earlier form.

**The Strength of the Relationship Between Variables**

It is important to emphasize that the existence of an empirical relationship is not a dichotomous state of affairs, in which $X$ and $Y$ are either related or not related to each other. Instead, empirical relationships are matters of degree. In fact, the clarity of the relationship between $Y$ and the independent variable(s) can, itself, be viewed as a phenomenon that varies along a continuum from a completely null state at one extreme (i.e., the variables are not related at all) to a completely deterministic state at the other (i.e., the variables are as strongly related as possible). The nature of these extremes can be defined very easily.

A null relationship exists when the conditional distributions of $Y$ are identical to the marginal distribution in all respects (except for the numbers of observations, which are $N_1$, $N_2$, …, $N_m$ for the respective conditional distributions and $N$ for the marginal distribution). In that case:

$$\mu_{Y|X_1} = \mu_{Y|X_2} = \ldots = \mu_{y|X_m} = \mu_Y$$

Now, it can be proven that the overall variance in the dependent variable is a function of the sum of the variance in the conditional $Y$ means and the mean of the conditional $Y$ variances. That is:

$$\sigma^2_Y = f(\text{var}(\mu_Y|X_i) + \text{mean}(\sigma^2_Y|X_i))$$

In the preceding expression “var” and “mean” are operators representing the variance or the mean of the expression within the parentheses immediately following the operator, itself. If the conditional means are all identical, then the first term on the right-hand side of the expression (i.e., $\text{var}(\mu_Y|X_i)$) must be equal to zero. That, in turn, implies that the conditional
variances are all constant and equal to the marginal variance of $Y$:

$$\sigma^2_Y|X_1 = \sigma^2_Y|X_2 = \ldots = \sigma^2_Y|X_m = \sigma^2_Y$$

Stated somewhat differently, if $Y$ is not related to $X$, then the conditional distributions are all identical replications of the marginal distribution.

Let us turn next to the opposite extreme on the continuum and consider the strongest possible relationship between $X$ and $Y$. In this situation, the conditional distributions “collapse” onto their respective means; within each of the conditional distributions, all of the observations take on the value $\mu_Y|X_i$. When this happens, there is no conditional variance at all:

$$\sigma^2_Y|X_1 = \sigma^2_Y|X_2 = \ldots = \sigma^2_Y|X_m = 0$$

Thus, if $Y$ is “perfectly” related to $X$, then each conditional distribution of $Y$ will actually consist of $N_i$ observations that take on the same $Y$ value, all of which are equal to $\mu_Y|X_i$.

Of course, these two extremes seldom arise (if they ever do) in real-life research situations. Empirical data never conform to the patterns we have described as either “null” or “perfect” relationships. But, they may approximate, or come closer to, one or the other of these patterns. And, that provides us with a useful strategy for defining the strength of the relationship between $X$ and $Y$. In order to illustrate this idea, let us make a simplifying assumption: We will assume that there are equal numbers of observations which possess each possible value of $X$. That is, $N_1 = N_2 = \ldots = N_m$. So, we will represent the number of observations at each $X_i$ as the single value, $N_i$. With this simplifying assumption, the expression for the marginal variance of $Y$ reduces to a simple sum:

$$\sigma^2_Y = \text{var}(\mu_Y|X_i) + \text{mean}(\sigma^2_Y|X_i)$$

Note that our simplifying assumption is not absolutely necessary. Even without it, the
right-hand side of the preceding equation would just be a weighted sum, with the weights
determined by the differing numbers of observations that fall at each $X_i$. For now, though,
we prefer to make our point with this simplest expression for $\sigma_Y^2$.

As we can see, the marginal variance of $Y$ is a sum of two components, the variance of
the conditional means and the mean of the conditional variances. If one of these components
is relatively large, then the other must be relatively small, since they sum to a constant
value (i.e., $\sigma_Y^2$). We will define the strength of the relationship between $X$ and $Y$ as the
degree to which the marginal $Y$ variance is composed of the variance in the conditional
means, rather than the mean of the conditional variances. The larger the first component
in the summed expression for the marginal variance (that is, $\text{var}(\mu_Y|X_i)$), compared to the
second component (that is, $\text{mean}(\sigma_Y^2|X_i)$), the stronger the relationship between $X$ and $Y$.
Conversely, the larger the second component relative to the first, the weaker the relationship
between $X$ and $Y$.

To explain the preceding definition a bit further, the variance in the conditional means
represents differences in the central locations of the conditional $Y$ distributions. And, it is
these very differences which we have used to define the relationship between the variables.
The greater the degree to which the overall variance in $Y$ is due to the shifting positions
of the conditional $Y$ distributions across $X$ values, the stronger the relationship between $X$
and $Y$. And, if we are willing to interpret this in causal terms, we might want to say that
$X$ has a greater impact on $Y$.

Let us consider the strength of the relationship from a forecasting perspective. Here, the
emphasis is on predicting the value of $Y$ for a specific observation, $i$. Start by assuming we
only have information about the marginal $Y$ distribution. In that case, our “best guess”
for $Y_i$ would be the mean, $\mu_Y$. And, our average prediction error in that case would be the
marginal variance, $\sigma_Y^2$. Now, assume we are given additional information, in the form of $i$’s
value on $X$, or $X_i$, along with information about the conditional $Y$ distributions (i.e., the
conditional means and conditional variances). How much does this additional information
improve our prediction of $Y_i$? Now, our best guess is the \textit{conditional} mean of $Y$, that is $\mu_{Y|X_i}$. And, our average prediction error would be the conditional variance, or $\sigma^2_{Y|X_i}$. The stronger the relationship between $X$ and $Y$, the smaller this conditional variance should be for any $X_i$. And, therefore, our prediction should be more accurate. Thus, the strength of the relationship between $X$ and $Y$ can be conceptualized as the improvement in our prediction of an observation’s $Y$ value when we know its value on $X$.

Figure 9 provides some illustrations of these ideas. Each panel uses box plots to represent the marginal distribution of a hypothetical $Y$, along with the conditional distributions across the values of a hypothetical $X$ variable (which takes on integer values from one to ten). The box plot for the marginal distribution is shown at the left side of each figure, while the box plots for the conditional distributions are shown within the interior of the respective panels. Now, box plots show the median and the inter-quartile range of a distribution, rather than the mean and the variance. But, in perfectly symmetric distributions, the median will be equal to the mean, and the interquartile range will be proportional to the standard deviation (which is, itself, the square root of the variance). The distributions in Figure 9 represent hypothetical data, with perfectly symmetric marginal and conditional distributions. Therefore, we can use the box plots to illustrate our ideas about relationships between variables.

Figure 9A illustrates a null relationship. Here, the ten conditional distributions are all identical to the marginal distribution. So, the conditional means do not vary, as indicated by the fact that the horizontal line segments within the boxes (which represent the central tendencies of the distributions) all fall at exactly the same height within the panel. And, the boxes in Figure 9A are all exactly the same height, showing that the conditional variances are all equal to the marginal variance. Of course, the latter implies that the mean of the conditional variances is also equal to the marginal variance. Stated differently, our average error in predicting an observation’s $Y$ value would be identical whether or not we know the observation’s value on $X$; the additional information provides no improvement whatsoever.
Figure 9B shows a perfect relationship. Here, the box plots for the conditional distributions are shown as horizontal line segments, because all observations with a common $X$ value, say $X_i$, have exactly the same $Y$ value. And, the latter is $\mu_{Y|X_i}$ (which is also equal to the conditional median). For all $X_i$, $\sigma^2_{Y|X_i} = 0$, so the boxes for the conditional distributions all have zero height. Here, we achieve perfect prediction of an observation’s $Y$ value, simply by knowing its $X$ value; the latter information reduces our prediction error to zero.

Figure 9C shows an intermediate situation. Here, we can see that there is a systematic shift in the positions of the conditional distributions across the $X_i$: The locations of the distributions move upward across the first seven $X$ values, but then move slightly downward for the last three $X$ values. So, the conditional means do exhibit some variance (as indicated by the differing positions of the horizontal bars within the conditional box plots), although its overall size relative to the marginal distribution of $Y$ is not very large. At the same time, notice that the heights of the conditional boxes are all somewhat smaller than the height of the marginal box. This shows that the conditional variances are smaller than the marginal variance— but, again, not by a large margin. These two features of the conditional distributions relative to the marginal distribution (i.e., the slight movement in their means and the slightly smaller variances) suggest that $X$ and $Y$ are weakly related to each other. Only a small portion of the marginal variance is composed of variance in the conditional means; most of it remains due to the mean of the conditional variances. (Note that the conditional variances are all equal in Figure 9C, so the mean is equal to the conditional variance of any of the ten subsets). From the prediction perspective, knowing $X_i$ will lead us to predict $Y$ values that are different from the overall $\mu_Y$, but there will still be a rather large amount of error in our prediction, on average.

Figure 9D shows another intermediate situation. But here, the shifts in the positions of the conditional box plots are larger than in the previous case, and the the heights of the conditional boxes are smaller. In other words, the conditional means vary quite a bit across the subgroups defined by the $X$ values, and the conditional variances are much smaller than
the marginal variance. Accordingly, this illustrates a relatively strong relationship between
X and Y. Here, much of the marginal Y variance is due to the variability in the conditional
Y means, leaving a fairly small portion that is composed of the mean from the conditional
variances. And, knowing $X_i$ would generate relatively small errors (on average) when we use
$\mu_Y|X_i$ as our prediction of an observation’s Y value.

Hopefully, inspecting the panels of Figure 9 will convince you that our formal definition
of a relationship between two variables corresponds perfectly with intuitive notions. In each
case, the existence of a relationship is signalled by the presence of a systematic pattern in
the differences across the conditional Y distributions, when the latter are compared to the
marginal Y distribution. And, the clarity of the pattern (which, itself, varies separately from
the basic existence of the pattern) corresponds to the strength of the relationship.

Note, too, that we can easily generalize our understanding of the strength of a relationship
to situations that involve more than one independent variable. In that case, the conditional
distributions would simply be defined by the distinct sets of values that occur across the
$k$ different $X_j$’s. Once again, we will make our simplifying assumption that every distinct
combination of $X_j$ values occurs for exactly the same number of observations (which we will
show as $N_i$). In that case, the marginal variance of $Y$ can be expressed as:

$$\sigma_Y^2 = \text{var}(\mu_Y|X_{1i}, X_{2i}, \ldots, X_{ki}) + \text{mean}(\sigma_Y^2|X_{1i}, X_{2i}, \ldots, X_{ki})$$

Here, the strength of the relationship between $Y$ and the full set of $X_j$’s is still defined as the
degree to which the marginal variance of $Y$ is composed of the variance in the conditional
means, or the first term on the right-hand side of the preceding equation.

**Models, Residuals, and Goodness of Fit**

Up to this point in our discussion of relationships between variables, we have talked about
expressing the variance in $Y$ using the variance in the conditional means calculated within
the subgroups of observations defined by the $X_i$’s. Let us introduce an important new term:
What we have been doing is using the conditional distributions to construct a *model* or the variance in \( Y \). A model can be defined as an abstract simplification of some phenomenon which is intended to represent the essential interesting structure in that phenomenon. Here, the “phenomenon” is the marginal variance in \( Y \). And, the model is composed of the differences across the conditional distributions; that is, we are using variability in these conditional distributions which occurs across the \( X_i \)'s to represent the overall value of \( \sigma_Y^2 \).

The specific elements used to achieve this representation are called the “parameters” of the model. In this case, the parameters are the conditional \( Y \) means, the \( \mu_{Y|X_i} \). To the extent that our model is an accurate representation of the phenomenon of interest, we should find that the variance in the conditional means approximates the value of \( \sigma_Y^2 \). If our model is *not* an accurate representation of the variance in \( Y \), then we will find that \( \text{var}(\mu_{Y|X_i}) < \sigma_Y^2 \); in other words, the variance in the conditional means does not comprise all of the marginal variance.

But, since we have already said that perfect relationships never occur with “real” data, it will always be the case that the variance of the conditional means *is* smaller than the marginal \( Y \) variance. So, what about the “left-over” part of \( \sigma_Y^2 \)? We will call this the “residual variance” in \( Y \), defined simply as the portion of \( Y \)'s variance that is not represented by our model. Within each of the subsets in our data defined by the \( X_i \)'s, the conditional variance represents the dispersion of the observations around the conditional mean. But, the latter represent the model parameters. Therefore, each conditional variance represents a part of \( Y \)'s variance that is not captured within our model. And, the overall amount of this left-over variance is aggregated into the average of the conditional variances, or mean(\( \sigma_Y^2 | X_i \)).

Now, models are artificial inventions of the researcher which may represent the phenomenon of interest with greater or lesser accuracy. And, with any particular model, it is highly desirable to have a measure that indicates how accurately that model depicts the relevant phenomenon. Ideally, we would like to convey this degree of accuracy with a single numeric value, which we will call a measure of “goodness of fit” for that model.
It is a straightforward task to develop a goodness-of-fit measure for our model of $Y$’s variance. Recall that this variance can be broken down into two additive components, $\text{var}(\mu_Y|X_i)$ and $\text{mean}(\sigma^2_Y|X_i)$. And, our model uses the former component to represent $\sigma^2_Y$, leaving the latter as the residual. Therefore, we can just define a goodness-of-fit measure (which we will call “GoF” for now) as the proportion of $Y$’s variance that is comprised of variance in the conditional means, as follows:

$$GoF = \frac{\text{var}(\mu_Y|X_i)}{\sigma^2_Y}$$

The value of $GoF$ can range from zero (which occurs when the conditional distributions do not differ from each other, and the conditional variances are equal to the marginal variance) to one (which occurs when the residual variance is zero, and the conditional means comprise all of the variance in $Y$). Generally speaking, we would prefer that $GoF$ be close to its upper bound of one, since that corresponds to a model which represents the variance in $Y$ with a high degree of accuracy.

Many analysts would say that the value of $GoF$ represents the proportion of variance in $Y$ that is “explained” by the model, and that the goodness-of-fit measure indicates a model’s “explanatory power.” While we will use this same terminology ourselves, we have to admit some uneasiness in doing so. The problem is that phrases like “explained by” and terms like “explanatory power” seem to suggest an inherently causal connection between the source of the model (in this case, the independent variable, $X$) and the phenomenon that is being represented (in this case, the variance in $Y$). While that may not be an unreasonable interpretation in some contexts, it is still important to recognize that two or more variables may be related to each other without any immediate causal connection between them. Nevertheless, such terminology finds common usage in the social science literature, so we will employ those terms as well. Doing so should cause no serious problems, as long as we all remember the caveats that are involved.
Table 3 shows the conditional means and conditional variances for the data that were used to produce the four panels of Figure 9. The marginal mean and marginal variance is the same for each panel. The table also shows the GoF value for the data in each panel. Here, we can see that the GoF for Figure 9A is 0.00, indicating that the model provides no information at all about the variance in \( Y \). The GoF for Figure 9B is 1.00, indicating that the variance in \( Y \) is depicted with complete accuracy by the conditional means. The GoF values for Figures 9C and 9D are xx and xx, respectively. These latter values suggest an important idea: Stronger relationships between variables correspond to models that fit the data more closely.

This latter idea completes a full circle and brings us back to our initial motivation for considering the concept of a relationship between variables. In empirical research contexts—regardless whether we are testing a theory or trying to achieve predictions of future events—we try to construct a model which accurately depicts the variance in our dependent variable of interest. A powerful model, in the sense of a structure which accounts for a high proportion of the variance in the dependent variable, reflects a strong empirical relationship between the latter and the independent variable (or variables). And, the latter is both a central empirical manifestation of a causal connection between the variables (which is, itself, the essential content of a substantive theory) and a central precondition for systematically accurate predictions (which is important for forecasting purposes). Thus, as practicing social scientists, we seek to construct powerful models that account for interesting variance in empirical variables. And, given the title of this book, you have probably already guessed that regression analysis is a tool for doing exactly that!

**Models and Parsimony**

Goodness of fit is a very desirable characteristic for any model of empirical data. But, it is definitely not the only criterion involved in model selection. A second desirable characteristic is parsimony. In other words, we would like our models to be as simple as possible. This general idea is embodied in principle commonly known as “Occam’s razor.” Given two
explanations that account equally well for the observed facts, the simpler explanation is to be preferred over the more complex explanation.

In the present context, “simplicity” or “parsimony” is defined in terms of the number of parameters used by a model to account for the variance in the dependent variable. So, the number of parameters in the model should be relatively small, compared to the size of the phenomenon being represented by that model. Note that we include the adverb “relatively” to indicate that there are probably no universal, objective, standards to indicate what “small” is. Instead, this must be judged on a context-specific basis, using criteria that are established by the researcher (or, perhaps, by the audience to which that researcher is trying to communicate his/her findings).

Here, the phenomenon of interest is the variance of $Y$, or $\sigma_Y^2$. This variance is a manifestation of dispersion across the $Y$ values in the entire population. Therefore, the size of the phenomenon to be explained is $N$, or the number of elements in the population, itself. What would be considered “small” relative to this value?

In our discussion about the strength of the relationship between $X$ and $Y$, we never said how many objects possessed each distinct value of $X$. We did, however, make the simplifying assumption that each $X_i$ occurred an equal number of times. Therefore, if there are $m$ distinct $X_i$’s, then $N = mN_i$. Now, our model uses the conditional means as parameters to represent $\sigma_Y^2$. There will be one conditional mean for each $X_i$. So, the total number of parameters is $m$. Is this a parsimonious model? The answer, of course, depends on $N_i$. If this value is large, indicating that there are many objects sharing each distinct $X_i$, then $N$ will be much larger than $m$. For example, in each panel of Figure 9, $N_i = 100$, which yields $N = 1000$, since there are $m = 10$ values of $X$. But, the number of parameters is also 10, since that is the number of conditional means. So, the number of model parameters is one-hundredth (i.e., 10/1000) the size of the phenomenon being explained. By any reasonable standard, this model is definitely parsimonious. If this parsimonious model also represents $\sigma_Y^2$ with a reasonably high degree of accuracy (as is the case for the data from Figures 9B and 9D),
then we might find this model to be useful for theory-testing and/or forecasting.

Let us consider a rather extreme counter-example which should help illustrate why parsimony is important, along with explanatory power, in assessing the utility of a model. What if we encountered a population in which there is only one observation at each distinct value of $X$? Here, $N_i = 1$ and $\mu Y|X_i$ is simply equal to $Y_i$. If we construct a model in which the variance of $Y$ is represented using the conditional means, we will achieve perfect goodness of fit. After all, with only one observation at each distinct $X_i$, it is impossible to have any dispersion around the conditional means. When aggregated across the $m$ values of $X$, the GoF will be 1.00. But, the number of parameters in this model will be equal to $N$, the total number of objects in the population, itself. Hence, the size of the model is equal to the size of the phenomenon we are trying to represent. In this situation, our model provides no simplification, and we might just as well stay with the original data. What is the problem with doing exactly that? Well, if $N$ is any larger than a tiny number, it will be almost impossible to discern any kind of systematic pattern within the population. And, as a result, the model provides us with no substantive benefit whatsoever.

Parsimony and goodness of fit are both desirable characteristics of a model. However, there is a tension between these two characteristics: Generally speaking, models with more parameters can represent a greater proportion of the variance in $Y$ than those with fewer parameters, and vice versa. So, these two characteristics seem to push us in opposite directions during the process of constructing models. For better or worse, this is a conflict that we will never be able to escape. And, this means that all assessments of parsimony and goodness of fit are necessarily context-specific and relative; there are simply no absolute standards that we can adopt or recommend. This is definitely a frustrating state of affairs! But, we believe that it is better to recognize it at the outset, and understand the implications of the tension between parsimony and fit, than to ignore the problem entirely.

**CONCLUSION**

In concluding this chapter, we now return to the question that we raised at its outset:
What is regression analysis? In light of the preceding discussion about empirical relationships and models, we respond by saying that regression analysis is a tool for constructing a model of the variance in some interesting dependent variable. It does so by providing a parsimonious description of the differences in the conditional distributions of the dependent variable across the values of one or more independent variables.

This answer is a bit more detailed than our earlier response (“Regression analysis is a tool for describing relationships among variables” from page 1). Nevertheless, it remains very general. The current answer says nothing about the specific nature of the model that we will create, nothing about the procedures that we will use to obtain values for the model parameters, and nothing about strategies for assessing model adequacy. Despite everything it seems to omit, however, we maintain that this answer really does capture the essential character of regression analysis in a succinct pair of sentences. And, the material presented in the remainder of this book simply elaborates upon this straightforward definition. Of course, in doing this, we will provide the information that is missing from our definition, in roughly the order that we mentioned these omissions.

Let us look ahead beyond the introductory material presented above and in Chapter 2, below. Section II (Chapters 3 through 7) will develop the basic elements of the regression model, while Section III (Chapters 8 through 11) demonstrate the flexibility of this model for dealing with various forms of data and structure that a researcher might encounter during the course of an empirical analysis. Sections IV (Chapters 12 through 15) and V (Chapters 16 through 18) deal with situations where interpretation of the model parameters are rendered more difficult. Most (but not all) of these situations arise when the basic assumptions of the regression model are not met in the data at hand. And, we will see that there are usually strategies which can be employed to deal with the resultant problematic aspects of the analysis, even if there is no way to “solve” most of the problems, themselves. Overall, our objective with this book is to familiarize readers with the technical aspects of regression analysis, while never losing sight of the basic fact that regression is a tool for answering
interesting substantive questions about data.
Figure 1.1: Producing a scatterplot from the data in Table 1.1.
Figure 1.2: A scatterplot showing data in which there is no apparent systematic structure.
Figure 1.3: Scatterplot showing systematic structure in bivariate data (this graphical display is identical to Figure 1.1D).
Figure 1.4: Scatterplot showing systematic structure in bivariate data, with line superimposed over the plotted data points.
**Figure 1.5:** Varying the degree to which the fitted “curve” (actually, a straight line) describe the data.

A. Points widely scattered around best-fitting line

![Graph A](image1)

B. Points fall closer to the best-fitting line

![Graph B](image2)

C. Points fall very close to the best-fitting line

![Graph C](image3)

D. Points fall exactly along the best-fitting line

![Graph D](image4)
Figure 1.8: Fitting an appropriate curve to bivariate data that exhibit a nonlinear pattern.