CHAPTER 1. LINEAR REGRESSION

Introduction

It is likely that you have thought about a variety of relationships between different variables. For example, on a personal level you might have wondered whether studying an additional hour will lead to a substantially better score on an upcoming exam. Or, will getting additional experience in a nonpaid internship increase the likelihood of obtaining a full-time position in an organization you wish to work for. Or, if you go to the gym an additional hour each week, will this lead to a substantial weight loss.

Relationships between variables are also at the heart of many academic disciplines. Political scientists may focus on the possible link between contributions to political leaders and decisions those leaders subsequently take. Economists may be interested in how mortgage interest rates affect the housing market. Marketing firms are likely interested in how different forms of advertising lead to improved sales of a product. Policy analysts are often interested in how a change in a policy, such as a new curriculum in elementary education, might lead to different outcomes, such as improved test scores. Public health researchers might be interested in determining how the incidence of cancer is related to the amount of red meat that is consumed. Others might be interested in how student standardized test scores are related to the marital status, education, and income of the parents, or how wages in metropolitan areas are related to the number of immigrants in the metropolitan area.

The point is that interest in understanding relationships is common and widespread. Researchers, both in the natural and social sciences, often want to delve much more deeply into the nature of those possible relationships. And when the variables of interest, such as contributions to political office seekers and votes on a particular issue, can be quantified, a very common method used for analyzing those relationships is linear regression analysis. Regression analysis is a statistical technique that provides a way of conveniently summarizing the relationship between a variable of interest and one or more variables that are anticipated to influence that variable.

This volume is about linear regression analysis, that is, analysis of cases in which the relationship between the variable to be explained and the other variable or variables can be summarized by a straight line. The volume is intended to provide the reader with a basic understanding of how regression analysis can be carried out, how the results from such analysis are interpreted, and the variety of ways in which regression analysis is used both in academic settings and in public and business arenas. The current chapter illustrates how a single variable can be used to explain variations in another variable, for example, the influence of mortgage interest rates on the number of new houses constructed. Chapter 2 shows how more complex relationships, in which a single variable is hypothesized to depend on two or more variables, can be estimated using regression analysis. For example, how do volume of traffic, speed, and weather conditions affect the number of accidents on highways? In most applications of regression analysis, researchers rely on data that constitute a sample drawn from a population. As is shown in Chapter 3, in these instances it is necessary to test hypotheses in order to generalize the findings to the population from which the sample was drawn. The final two chapters expand on the discussion of regression analysis. Chapter 4 focuses on the data used and Chapter 5 on a variety of problems and issues researchers face when using this technique. Throughout the volume we keep the discussion as simple as possible and provide examples to illustrate how regression analysis is applied in a variety of disciplines. Our objective is to give the reader a solid but basic understanding of linear regression analysis, not to make the reader an expert. Thus many more complex statistical issues are not covered in this book. Readers who wish a more in-depth coverage are referred to the suggested readings provided in Appendix D.

Hypothesized Relationships

The two statements, “The more a political candidate spends on advertising, the larger the percentage of the vote the candidate will receive” and “Mary is taller than Jane,” express different types of relationships. The first statement implies that the percentage of the vote that a candidate receives is a function of, or is caused by, the amount of advertising, while in the second statement, no causality is implied. More precisely, the former expresses a *causal* or *functional* relationship while the latter does not. A functional relationship is a statement (often expressed in the form of an equation) of how one variable, called the *dependent* variable, depends on one or more other variables, called *independent* or *explanatory* variables.[1] In the example, the share of the vote a candidate receives is dependent on (is a function of) the amount spent on advertising. Another independent variable that might be included in the analysis is the number of prior years in office, in which case the functional relationship would be stated as, “The candidate’s share of the vote depends on the amount of advertising as well as the candidate’s prior years in office.”

Researchers are often interested in testing the validity or falsity of hypothesized functional relationships, called *hypotheses*[2] or *theories.* We show in Chapter 3 how linear regression is used to test such hypotheses. But first we explain how a regression equation is estimated.

Linear regression analysis is applicable to a vast array of subject matter. Consider the following situations in which regression analysis has been employed: a study of the effect of polluting industries on mortality rates in Chinese cities (Hanlon and Tian 2015), a study of how proximity to fast-food restaurants relates to the percentage of ninth graders who are obese (Currie, Della Vigne, Moretti, and Pathania 2010), a study of the relationship between income tax rebates and sales of hybrid electric vehicles (Chandra, Gulati, and Kandikar 2010), a study of the effect of changes in cigarette prices on smoking among smokers of different smoking intensities (Cavazos-Rehg et al. 2014), and a study showing the effect of severe drops in temperature on the number of trials for the crime of witchcraft in 16th and 17th century Europe (Oster 2004). All of these examples are cases in which the application of regression analysis was useful, although the application was not always as straightforward as the example to which we now turn.

A Numerical Example

To facilitate the discussion of linear regression analysis, the following food consumption example will be referred to throughout the book. Suppose one were asked to investigate by how much a typical family's food expenditure increases as a result of an increase in its income. While most would agree that there is a relationship between income and the amount spent on food, the example is in fact an investigation of an economic theory. The theory suggests that the expenditure on food is a function of family income;[3] that is, $C = f\left(I\right)$, read "*C* is a function of *I*," where *C* (the dependent variable) refers to the expenditure on food, and *I* (the independent variable, sometimes called the regressor) denotes income. Throughout the book we will refer to the theory that *C* increases as *I* increases as the hypothesis.[4]

The investigation of the relationship between *C* and *I* allows for both testing the theory that *C* increases as a result of increases in *I* and obtaining an estimate of how much food consumption changes as income changes. One can therefore consider the investigation as an analysis of two related questions: (l) Does spending on food increase when a family's income increases? (2) By how much does spending on food change when income increases or decreases? In this chapter we explain how a linear relationship between the two variables is estimated in order to provide descriptive answers to these questions, although, as will be seen in Chapter 3, these questions cannot be answered with certainty.

A common strategy for answering questions such as these is to observe income and food consumption differences among a number of families and note how differences in food consumption are related to differences in income. Here we employ the hypothetical data given in columns l and 2 of Table 1.1 to answer this question. The data represent annual income and food consumption information from a sample of 50 families in the United

Table 1.1 Food Consumption, Family Income, and Family Size Data

(1) *Food Consumption*	*(2)* *Income*	*(3)* *Family Size*	*(4)* *Has a Garden*
$780	$24,000	1	NO
1,612	20,000	1	NO
1,621	37,436	1	NO
1,820	36,600	2	YES
2,444	10,164	1	YES
3,120	2,500	1	NO
3,952	29,000	1	YES
4,056	40,000	1	NO
4,160	30,154	1	NO
4,160	34,000	1	YES
4,300	46,868	1	NO
4,420	15,000	1	NO
5,200	36,400	2	YES
5,200	25,214	2	YES
6,100	21,400	2	YES
6,240	68,620	2	YES
6,587	1,200	3	NO
7,020	40,000	2	NO
7,040	52,000	1	NO
7,540	31,100	2	NO
7,600	107,602	4	NO
8,060	134,000	2	NO
8,632	59,800	3	NO
8,800	68,000	4	NO
8,812	80,210	2	NO
8,840	67,000	1	NO
9,100	50,000	6	NO
9,150	53,420	1	NO
9,360	55,000	1	NO
9,658	65,000	1	NO
9,660	66,000	2	YES
9,880	28,912	3	YES
10,192	100,000	1	YES
10,296	50,356	4	YES

(1) *Food Consumption*	*(2)* *Income*	*(3)* *Family Size*	*(4)* *Has a Garden*
10,400	45,000	4	NO
11,263	168,000	4	NO
11,700	110,200	4	NO
11,960	75,000	3	NO
12,036	150,200	1	YES
12,064	44,746	2	NO
12,240	171,170	2	NO
12,652	170,000	5	NO
13,260	27,000	2	NO
14,377	132,543	2	YES
14,731	192,220	2	NO
15,300	141,323	4	NO
16,584	84,059	2	NO
16,870	176,915	5	NO
18,776	189,654	5	NO
20,132	151,100	3	NO

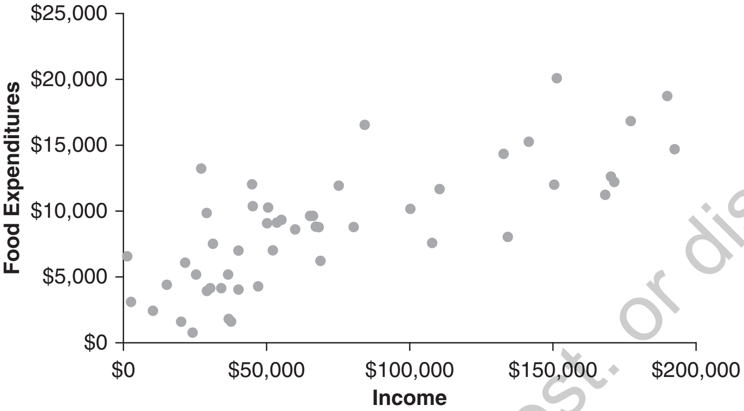
Source: Hypothetical data

States for one year. Assume that this sample was chosen randomly from the population of all families in the United States.[5] The associated levels of these two variables have been plotted as the 50 points in Figure 1.1. We are going to use this sample to draw inferences about how income affects food consumption for the population of families.

Casual observation of the points in Figure 1.1 suggests that *C* increases as *I* increases. However, the magnitude by which *C* increases as *I* increases for the 50 families is not obvious. For this reason the presentation of data in tabular or graphical form is not by itself a particularly useful format from which to draw inferences. These formats are even less desirable as the number of observations and variables increases. Thus we seek a means of summarizing or organizing the data in a more useful manner.

Any functional relationship can conveniently be expressed as a mathematical equation. If one can determine the equation for the relationship between *C* and *I*, one can use this equation as a means of summarizing the data. Since an equation is defined by its form and the values of its parameters, the investigation of the relationship between *C* and *I* entails learning something from the data about the form and parameters of the equation.

Figure 1.1 Scatter Diagram of Family Income and Food Consumption



The economic theory that suggests that *C* is a function of *I* does not indicate the form of the relationship between *C* and *I*. That is, it is not known whether the equation is of a linear or some other, more complex form.[6] In some problems the general form of the equation is suggested by the theory, but since this is not so in the food expenditure problem, it is necessary to specify a particular form. We shall assume that the form of the equation for our problem is that of a straight line, which is the simplest and most commonly used functional form.[7] (A review of the algebraic expression for a straight line is given in Box 1.1.)

Box 1.1 Algebraic Expression for a Straight Line

As you may remember from algebra, a straight line relating two variables *X* and *Y,* with *Y* considered a function of *X,* can be expressed using the formula $Y = a + bX$, where *a* and *b* are numbers; *a* is the intercept, which is the value of *Y* when *X* is zero, and *b* is the slope, which measures the change in *Y* associated with a unit increase or decrease in *X*. For example if $Y = 2.5 + 0.7X$, a two-dimensional graph with *Y* on the vertical axis and *X* on the horizontal axis would show the line passing through the vertical axis at $Y = 2.5$ (since *X is* zero at that point) and with a slope of 0.7, which means that for any one-unit increase in *X* there is a 0.7-unit increase in *Y.*

Given this assumption, one can express the functional relationship that exists between C and I for all U.S. families as

$$C = \alpha + \beta I \qquad [1.1]$$

where α (the Greek letter alpha) and β (the Greek letter beta) are the unknown parameters assumed to hold for the population of US families, and are referred to as the *population parameters.*

Given the assumption that the form of the equation of the possible relationship between C and I can be represented by a straight line, what remains is to estimate the values of the population parameters of the equation using our sample of 50 families. The two questions posed earlier refer to the value of the slope—that is, the value of β. The first question asks whether β is greater than zero, while the second asks the value of β. By obtaining an estimate of the value of β, a statement can be made as to the effect of changes in income on the amount spent on food for the 50 families in our sample. As shown in Chapter 3, inferences can be drawn from this estimate of β about the behavior of all families in the population.

Before proceeding, it is important to note the following. The actual or "true" form of the relationship between I and C is not known. We have simply assumed a particular form for the relationship in order to summarize the data in Figure 1.1. Further, we do not know the values of the population parameters of the assumed linear relationship between C and I. The task is to obtain estimates of the values of α and β. We will denote these estimates as a and b.[8]

Estimating a Linear Relationship

The question that may come to mind at this point is, "How can it be stated that income and food consumption are related by a precise linear equation when the data points in Figure 1.1 clearly do not lie on a straight line?" The answer comprises three parts. First, the assumption that a straight line is a good summary of the data points does not imply that C and I are related in precisely this manner for every family. Second, the hypothesis is based on the implicit assumption that only income and food consumption differ between these families. However, other things, such as family size and tastes, are not likely to be the same for all families and thus affect the amount spent on food. Third, there is randomness in people's behavior; that is, an individual or family, for no apparent reason, may buy more or less food than some other family that appears to be in exactly the same situation with regard to income, taste, and the like. Thus one would not expect the data points to lie consistently on a straight line even if the line did represent the average response to changes in income.

More formally the regression equation is expressed as

$$C_i = \alpha + \beta I_i + \in_i \qquad [1.2]$$

In this form we say that food consumption of the *ith* family, C_i, depends on its income, I_i. The term ϵ_i is referred to as the *error term*, and captures the fact that given the values of α and β, the equation will not exactly predict a family's food consumption given its income for reasons discussed in the previous paragraph.

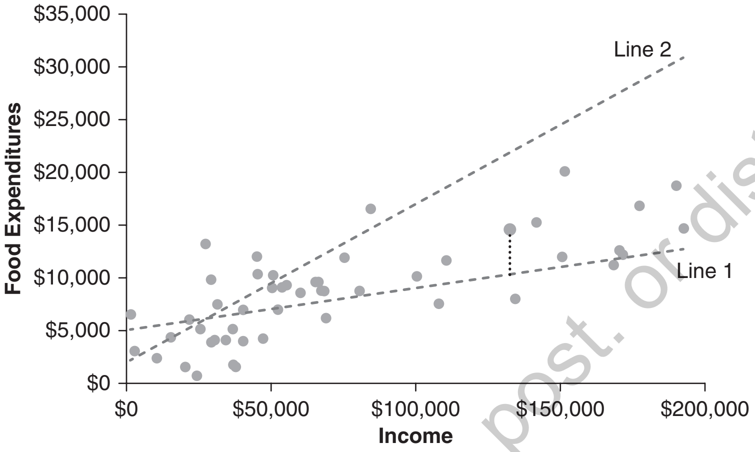
As noted previously, from the data points in Figure 1.1 it is not obvious how much C increases as I increases; that is, it is uncertain what the slope of the line summarizing the data points should be. To see this, consider the two solid lines that have been arbitrarily drawn through the points in Figure 1.2. Line 1 has the equation $C = 5000 + 0.04I$, and line 2 has the equation $C = 2000 + 0.15I$. Which of these two lines is the better estimate of how food consumption changes as income changes? This is the same as asking which of the two equations is better at summarizing the relationship between C and I found in Table 1.1. More generally, which line among all the straight lines that it is possible to draw through the points in Figure 1.2 is the "best" in terms of summarizing the relationship between C and I? Regression analysis, in essence, provides a procedure for determining the *regression line*, which is the best straight line (or linear) approximation of the relationship between C and I. This procedure is equivalent to finding particular values for the slope and intercept.

An intuitive idea of what is meant by the process of finding a linear approximation of the relationship between the independent and dependent variables can be obtained by taking a string or ruler and trying to "fit" the points in Figure 1.1 to a line. Move the string up or down, or rotate it until it takes on the general tendency of the points in the graph.

What property should this best-fitting line possess? If asked to select which of the two solid lines in Figure 1.2 is better at summarizing (estimating) the relationship between income and food consumption, one would undoubtedly choose line l, because it is "closer" to the points than line 2. (This is not to imply that line l is the regression line.)

Closeness or distance can be measured in different ways. Two possible measures are the vertical distance and the horizontal distance between the observed points and a line. In the normal case, where the dependent variable is plotted along the vertical axis, distance is measured vertically as the differences between the observed points and a fitted line. This is shown in Figure 1.2, where the vertical dotted line drawn from the data point to line l measures the distance between the observed data point and the line. In this

Figure 1.2 Two Possible Summaries of the Income-Consumption Relationship



case distance is measured in dollars of consumption, not in feet or inches. The choice of the vertical distance stems from the theory stating that the value of *C* depends on the value of *I*. Thus, for a particular value of income, it is desired that the regression line be chosen so as to predict a value of food consumption that is as close as possible to the value of food consumption observed at that income level.

The regression line cannot minimize the distance for all points simultaneously. In Figure 1.2 it can be seen that some points are closer to line 1, while others are closer to line 2. Thus a method of averaging or summing up all these distances is needed to obtain the best fitting line.

Although several methods exist for summing these distances, the most common method in regression analysis is to find the sum of the squared values of the vertical distances. This is expressed as

$$\Sigma_{i=1}^{n}(C_i - \hat{C}_i)^2 \qquad [1.3]$$

where C_i is the actual value of *C* for the ith family, $\hat{C}_i$ (read "C hat sub i") is the value of *C* for the *ith* family that would be estimated by the regression line and *n* is the number of observations over which the expression is summed.[9]

Least Squares Regression

In the most common form of regression analysis, the line that is chosen is the one that minimizes

$$\Sigma_{i=1}^{n}(C_i - \hat{C}_i)^2 \qquad [1.4]$$

which is called the *sum of the squared errors*, frequently denoted *SSE*.[10] For each observation, the distance between the observed and the predicted level of consumption can be thought of as an error, since the observed level of consumption is not likely to be predicted exactly but is missed by some amount $(C_i - \hat{C}_i)$. As noted above, this error may be due, for example, to randomness in behavior or other factors such as differences in family size. Because the squares of the errors are minimized, the term *least squares regression analysis* is used, with the estimation technique commonly referred to as *ordinary least squares* (OLS).

The reasons for selecting the sum of the squared errors lie in statistical theory that is beyond the scope of this book. However, an intuitive rationale for its selection can be presented. If the errors were not squared, distances above the line would be canceled by distances below the line. Thus it would be possible to have several lines, all of which minimized the sum of the nonsquared errors.[11] It is implicit that closeness is good, while remoteness is bad. It can also be argued that the undesirability of remoteness increases more than in proportion to the error. Thus, for example, an error of four dollars is considered more than twice as bad as an error of two dollars. One way of taking this into account is to weight larger errors more than smaller errors, so that in the process of minimizing it is more important to reduce larger errors. Squaring errors is one means of weighting them.

Let *a* and *b* represent the estimated values of *α* and *β* for the still unknown regression line. $\hat{C}_i$ can be expressed as $\hat{C}_i = a + bI_i$. Substituting $a + bI_i$ for $\hat{C}_i$ in expression 1.4, the expression for *SSE* can be rewritten as

$$SSE = \Sigma_{i=1}^{n}\left(C_i - a - bI_i\right)^2 \qquad [1.5]$$

Note that the term in parentheses in equation 1.5 is the error term, that is, an estimate of $\in_i$, from equation 1.2.

Expressions for *a* and *b* can be found that minimize the value of equation 1.5 and hence give the least squares estimates of *α* and *β*, which in turn define the regression line. (See Appendix A for the derivation of the formulas using the calculus.)

For the given set of data in Table 1, the *a* and *b* that minimize equation 1.5 are $a = 4{,}155.21$ and $b = +0.064$. (Statistical packages, which are readily available, generate the values of *a* and *b*. For purposes of completeness, Appendix A shows how the actual values of *a* and *b* are calculated.) Therefore, the least squares line, which is drawn in Figure 1.3, has the equation

$$C = 4{,}155.21 + 0.064I \qquad [1.6]$$

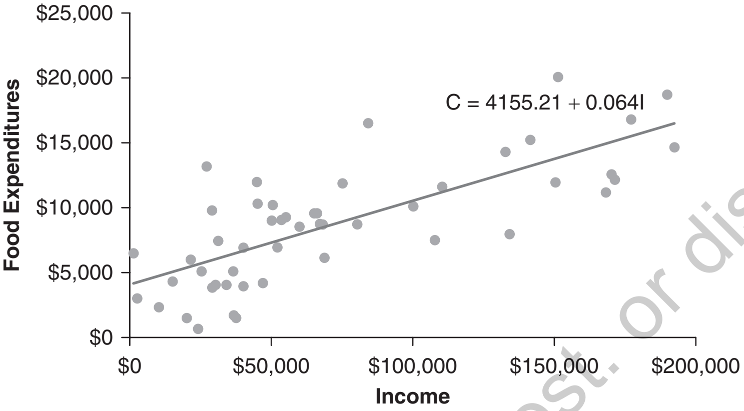
These results mean, for example, that the estimate of consumption for a family whose annual income is \$10,000 is \$4,795.21—that is, \$4,795.21 = \$4,155.21 + 0.064(\$10,000). Remember, this is an estimate of *C* and not necessarily the amount one would observe for a specific family with an income of \$10,000. The value of *a*, \$4,155.21, is the estimated food consumption for a family with zero income. The value of *b*, 0.064, implies that for this sample, each dollar change in family income results in a change of \$0.064 in food consumption in the same direction (note the positive sign for *b*). When interpreting the results of a regression analysis, it is important to keep in mind the unit of measure for the variables used in the equation (see Box 1.2).

Box 1.2 Importance of the Units of Measure

The estimate of the slope coefficient β is interpreted as the change in the dependent variable associated with a one-unit change in the independent variable. In our food example, both income and the amount spent on food are measured in dollars, so a one-unit change is one dollar. But what if each variable had been measured in thousands of dollars? In that case, observations for the first family in Table 1.1 would be 0.780 (thousands of dollars spent on food) and 24 (thousands of dollars income), and the regression result would be $C = 4.15521 + 0.064I$. Here the intercept represents the same 4.15521 thousands of dollars and for each thousand-dollar increase in income (a one-unit increase) there would be an associated 0.064 thousand (in other words, \$64) increase in food consumption. This of course is a 6.4 cent increase in food consumption per dollar increase in income.

But what if income were measured in thousands while *C* continued to be shown in dollars? Then the resulting equation would be $C = 4155.21 + 64I$. Again this would mean that for each additional one thousand dollars in income (i.e., a one-unit increase in *I*), there would be an associated \$64 increase in food consumption. The "story" about the relationship between *C* and *I* remains the same in spite of the different sizes of the coefficients. The important lesson to keep in mind is to be aware of the units of measure any time you interpret linear regression results.

Figure 1.3 "Best Fitting" Regression Line



These conclusions, of course, hold only for this particular sample. When the least squared technique is applied to additional samples of consumers, one would obtain additional (generally different) estimates of α and β.

It is important to point out that regression analysis does not prove causation. Our estimate of β is consistent with the theory that an increase in income causes an increase in food consumption. However, it does not prove causation. Note that we could have reversed the equation, making I depend on C, and argued that higher food consumption makes for healthier and more productive workers who thus have higher incomes. Since I and C increase together, this alternative relationship would also be supported. It would take some alternative experiment or test to determine the direction of the causation. Our estimate of β, however, is not consistent with the theory that food consumption decreases with increases in income.[12]

Note that linear regression requires that the regression equation relating the dependent and independent variables be linear, that is, a straight line. However, the equation relating Y and X does not have to be linear, so long as the estimated regression equation is linear. For example, suppose that we believe that the relationship between Y and X is given by $Y^2 = \alpha + \beta X^2$. Thus, Y and X are not related by a straight line, but the relationship between the variables Y^2 and X^2 is linear, so one can estimate a linear regression using Y^2 and X^2. We return to this topic in Chapter 4.

Examples

Before proceeding, three examples are presented to illustrate how regression analysis is used. Note that these examples have been selected in order to give the reader some idea of the variety of ways in which linear regression has been utilized in published research. They also represent only a portion of more extensive research that is included in the original articles; reading the entire article will provide much more information on the research that was conducted.

Example 1.1—Change in Education Performance

Goldin and Katz (2008, 346) examine the rise in education levels in the United States during the 20th century and its relationship to economic development. As part of their analysis, they explore how increased educational performance during the last two-thirds of the 20th century differed across states. In particular, they ask if states that had relatively low high school graduation rates in 1938 also had relatively low educational performance levels at the end of the century. To explore this question, they estimated a regression in which the independent variable is the state's high school graduation rate in 1938 and the dependent variable is an index of educational performance for the 1990s. The index averages several National Assessment of Education Progress (NAEP) scores, Scholastic Aptitude Test (SAT) scores, and a measure of the high school dropout rate. The regression is estimated for the 48 states that comprised the United States in 1938. The resulting regression is

$$EPI = -2.02 + 4.09HSG \qquad [1.7]$$

where *EPI* is the educational performance index and *HSG* is the high school graduation rate in 1938. The positive coefficient on *HSG* implies that, on average, states with high (low) high school graduation rates in 1938 had a high (low) educational performance index in the 1990s. In other words, performance in the 1990s is positively related to high school graduation rates in 1938, consistent with the idea that relative performance of states did not change greatly overtime.

Example 1.2—Women in Films and Box Office Receipts

It has been noted that female actresses are underrepresented in films. To explore a possible explanation as to why this might be the case, Lindner, Lindquist, and Arnold (2015) examine whether the presence of at least two women playing important roles in a film results in lower box office receipts. If this is the case, the result would suggest that the underrepresentation, at least in part, is due to lower public interest in such films. The authors estimate a regression in which the dependent variable is box office receipts and the explanatory variable is a measure of the presence of women in the film. They use data for 964 films for the period 2000–2009, and obtain the following regression equation:

$$R = 84.843 - 11.356F \qquad [1.8]$$

where R is the box office receipts (in millions of dollars) of a film and F is a measure of the gender representation in the movie. The negative coefficient of −11.356 implies that the presence of important roles by females in movies is associated with lower box office receipts, consistent with the premise that the public is less interested in seeing movies that feature females.

Example 1.3—Using Measures of Extremities to Determine Height

Forensic scientists can face the challenging task of determining the identity of individuals from commingled human remains. One variable that can be of use in determining identity is the approximate height of the presumed deceased. Linear regression analysis is one method that has been used to infer the height of the deceased when only measures of extremities, such as hands or feet, are available. For example, Krishan Kanchan, and Sharma (2012) obtained the following result when the height (HT) of 123 females living in Himachal Pradesh State in India was regressed on the length of their feet (FT), where each variable was measured in centimeters:

$$HT = 74.820 + 3.579FT \qquad [1.9]$$

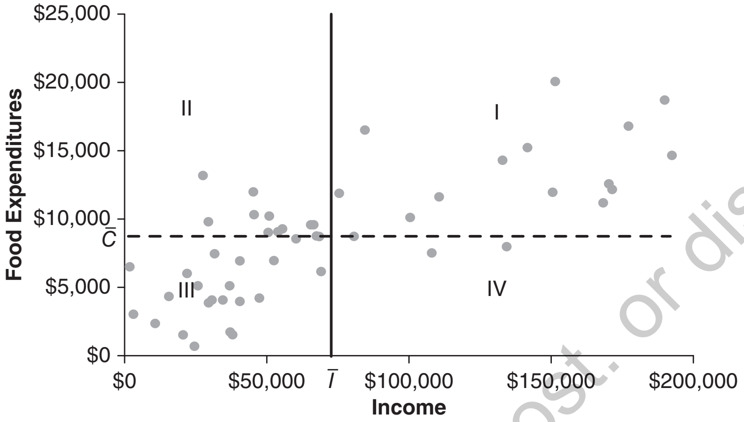
The results imply that each additional centimeter of foot length is associated with 3.579 cm. of additional height.

The Linear Correlation Coefficient

In the first part of this chapter, we demonstrated how regression analysis can be used to summarize the relationship between a dependent and an independent variable. We turn now to an explanation of descriptive statistics designed to evaluate (1) the degree of association between variables and (2) how well the independent variable has explained the dependent variable.

The *correlation coefficient* measures the degree of linear association between two variables.[13] To understand what statisticians mean by linear

Figure 1.4 Linear Correlation Analysis: The Food Expenditure Problem



association, consider Figure 1.4, which has the same 50 points as Figure 1.1. The average (or mean) level of food consumption is represented by the horizontal dotted line, while the vertical solid line represents the mean level of income. The two lines divide the figure into the four quadrants denoted by Roman numerals. Levels of *C* that are greater than the average of $8,795.14 lie above the dashed line in quadrants I and II, while less than average levels lie below, in quadrants III and IV. Similarly, income levels greater than the average lie to the right of $72,321.72 in quadrants I and IV, while those less than average lie to the left in quadrants II and III.

Figure 1.4 demonstrates that a majority of the points in the sample lie in quadrants I and III. Because of this pattern, the variables *C* and *I* are said to be *positively correlated.* Put differently, *C* and *I* are said to be positively correlated when *C*'s above (below) the mean value of food consumption, denoted $\bar{C}$, are associated with *I*'s above (below) the mean value of income, denoted $\bar{I}$. On the other hand, if the *C*'s below $\bar{C}$ had been associated with the *I*'s above $\bar{I}$ (and vice versa), one would have said that the variables were *negatively correlated.* The reader should be able to demonstrate that in this case the data points would have been clustered in quadrants II and IV. Other possibilities exist: If the data points had been spread fairly evenly throughout the four quadrants or in just quadrants II and III or just III and IV, one would have said that *C* and *I* were *uncorrelated.*

The particular descriptive statistic that measures the degree of linear association between two variables is called the *correlation coefficient* and is denoted *r*.[14] Although we don't provide the proof, *r* always lies between the values of −1 and +l ($-1.0 \le r \le +1.0$). When there is little linear association between two variables (when two variables are relatively unrelated), r is close to zero. In the presence of strong correlation, r is close to 1 (+1 for positive correlation, −1 for negative correlation).

Although the correlation coefficient is 0.756 (a positive number) for the food example, where it was hypothesized that changes in income caused changes in food expenditures, the presence of either positive or negative correlation does not necessarily indicate causality. In particular, because the correlation coefficient measures only the degree of association between two variables, it might reflect a cause-and-effect relationship; however there are other reasons besides causality that can influence the size of the coefficient. Variables may also appear correlated if both variables affect each other, if the two variables are both related to a third variable, or if the variables are systematically associated by coincidence.

An example of the first reason that both variables might affect each other is that IQ scores and student achievement scores are likely to be positively correlated. Although it seems reasonable that IQ influences achievement, many educators believe that this is only part of the story. Indeed, it seems likely that the IQ measure also reflects the level of achievement. An example of the second reason, that is, that the variables are related to a third variable, is the positive correlation that exists across cities between the number of churches and the number of bars. Although churches may spring up in response to bars (or bars in response to churches), the positive association most likely results because both variables are related to some other variable, such as population. A good example of the last reason, that the variables are related by coincidence, is the positive correlation of 0.943 found between the number of letters in the names of the teams in the Central Division of the National Baseball League and the number of wins during the 2014 regular season.[15]

The Coefficient of Determination

Recall that for any problem, the regression line is defined to be the line lying closest to the data points (closest in the sense that the line minimizes the sum of the squared errors term). Often, for comparative purposes, it is useful to know just how close is "close"; in other words, it is helpful to be able to evaluate what is referred to as the *goodness of fit* of the regression line.

An intuitive feeling for what is meant by goodness of fit is given in Figure 1.5, in which two distinct sets of data points have been plotted along with the two lines that minimize the sum of the squared errors. The two regression lines have the same values for a and b. The data points in panel

A of Figure 1.5 are clearly closer to the regression line than the data points in panel B.

The measure of relative closeness used by statisticians for evaluating goodness of fit is called the *coefficient of determination.* Because of its relationship to the correlation coefficient, this measure is generally referred to as the R^2. (The coefficient of determination is actually the square of the correlation coefficient, and is commonly referred to as “R squared.”) The R^2 statistic measures closeness as the percentage of total variation in the dependent variable explained by the regression line. Formally, the measure is defined as

$$R^2 = \Sigma_{i=1}^{n}(\hat{C}_I - \overline{C})^2 / \Sigma_{i=1}^{n}(\mathrm{C}_I - \overline{C})^2 \qquad [1.10]$$

To measure variation in a family’s food consumption, we want some common base from which to measure differences in *C*. To the extent that families consume more or less than the mean food consumption, denoted $\overline{C}$, there is variation in food consumption. Thus we use $\overline{C}$ as the base for measuring variations in *C* between families.

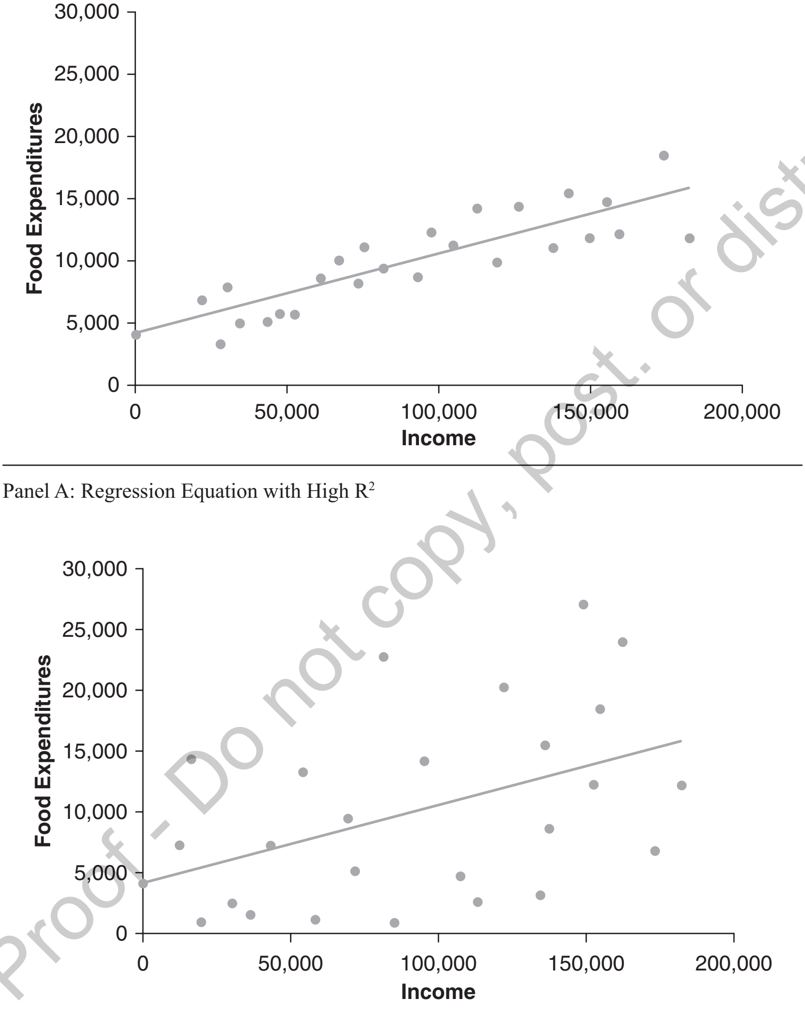
The denominator of equation 1.10 is a measure of the total variation in the dependent variable about its mean value $\overline{C}$. For example, consider a household with an income of $21,400 and observed consumption of $6,100 (the 15th observation shown in Table 1.1). Since the mean value of consumption is $8,795.14, the observed variation of *C* from the mean is −$2,695.14 for this observation (−$2,695.14 = 6,100 − 8,795.14). So that negative variations do not cancel positive variations, the individual variations are squared before they are summed.

The numerator of equation 1.10 is a measure of the total variation explained by the regression line. For example, from regression equation 1.6, it follows that the best estimate of food consumption for the family with an income of $21,400 is $5,524.81 ($5,524.81 = 4,155.21 + 0.064($21,400). Since this is −$3,270.33 from the mean (−$3,270.33 = $5,524.81 − $8,795.14), it is said that −$3,270.33 is the variation explained by the regression line for this observation. The total explained variation is found by summing the squares of these variations for the entire sample.

For the food expenditure problem, the value of the R^2 is 0.571, and one can say that the regression line explains 57.1% of the total variation in food expenditures. Stated somewhat differently, it can be said that 57.1% of the variation (about the mean) in the dependent variable has been explained by (or is attributable to) variation (about the mean) in the independent variable.

Notice that if the data points were all to lie directly on the regression line, the observed values of the dependent variable would be equal to the

Figure 1.5 Comparison of Goodness of Fit for Two Regression Lines



Panel A: Regression Equation with High R^2

Panel B: Regression Equation with Low R^2

predicted values, and the R^2 would be equal to l. As the independent variable explains less and less of the variation in the dependent variable, the value of R^2 falls toward zero. Hence, as would be expected, the R^2 for the data in panel A of Figure 1.5, 0.741, is greater than that for the data in panel B of Figure 1.5, 0.209.

Regression and Correlation

It is important to note that linear regression, the correlation coefficient, and the coefficient of determination are all related but that they provide different amounts of information and are based on different assumptions. First, as indicated previously, the coefficient of determination is simply the square of the correlation coefficient. An examination of Figure 1.4 should also convince the reader that if two variables are positively (negatively) correlated, the regression coefficient, that is, *b*, will have a positive (negative) sign.[16]

While this general relationship between *r* and *b* will always hold, one might ask if one of these two measures provides more information than the other. The answer is that the regression coefficient is more informative, since it indicates by how much and in what direction the dependent variable changes as the independent variable changes, whereas the correlation coefficient indicates only whether or not the two variables move in the same or opposite directions and the degree of linear association. This additional information from regression is obtained, however, at the cost of a more restrictive assumption—namely, that the dependent variable is a function of the independent variable. It is not necessary to designate which is the dependent and which the independent variable when a correlation coefficient is obtained.

Summary

Linear regression analysis provides a method for summarizing how one variable, referred to as the independent variable, explains variation in another, referred to as the dependent variable. In simple linear regression analysis, the relationship takes the form of a straight line defined by the slope coefficient and the intercept coefficient. The specific line chosen among all possible lines to summarize the relationship between the two variables is the one that minimizes the sum of the squared errors. The coefficient of determination provides a measure of the goodness of fit between the regression line and the data used to estimate the regression line. In the next chapter we examine how the same technique can be used to explore the relationship between a dependent variable and several independent variables.