
STATISTICAL METHODS FOR THE SOCIAL SCIENCES

Fifth Edition

Alan Agresti

University of Florida

LINEAR REGRESSION AND CORRELATION

Chapter 9

CHAPTER OUTLINE

- 9.1 Linear Relationships
- 9.2 Least Squares
Prediction Equation
- 9.3 The Linear
Regression Model
- 9.4 Measuring Linear
Association—The
Correlation
- 9.5 Inferences
for the Slope
and Correlation
- 9.6 Model Assumptions
and Violations
- 9.7 Chapter Summary

Chapter 8 presented methods for analyzing association between categorical response and explanatory variables. This chapter presents methods for analyzing association between quantitative response and explanatory variables. The analyses are collectively called a **regression analysis**.

Example 9.1 Regression Analysis for Crime Indices Table 9.1 shows data from *Statistical Abstract of the United States* for the 50 states and the District of Columbia (D.C.) on

- Murder rate: The number of murders per 100,000 people in the population.
- Violent crime rate: The number of murders, forcible rapes, robberies, and aggravated assaults per 100,000 people in the population.
- Percentage of the population with income below the poverty level.
- Percentage of families headed by a single parent.

For these quantitative variables, violent crime rate and murder rate are natural response variables. We'll treat the poverty rate and percentage of single-parent families as explanatory variables for these responses as we present regression methods in this chapter. The text website contains this data file, called **Crime2**, as well as a data file **Crime** that contains 2013 violent crime and murder rates already analyzed in Chapter 3. ■

We present three different, but related, aspects of regression analysis:

1. We investigate *whether an association exists* between the variables by testing the hypothesis of statistical independence.
2. We study the *strength of their association* using the *correlation* measure of association.
3. We estimate a *regression equation* that predicts the value of the response variable from the value of the explanatory variable.

9.1 Linear Relationships

We let y denote the *response* variable and let x denote the *explanatory* variable. We analyze how values of y tend to change from one subset of the population to another, as defined by values of x . For categorical variables, we did this by comparing the conditional distributions of y at the various categories of x , in a contingency table. For quantitative variables, a mathematical formula describes how the conditional distribution of y (such as y = crime rate) varies according to the value of x (such as x = percentage below the poverty level). Does the crime rate tend to be higher for states that have higher poverty rates?

TABLE 9.1: Statewide Data (from Crime2 Data File at the Text Website) Used to Illustrate Regression Analyses

State	Violent Crime	Murder Rate	Poverty Rate	Single Parent	State	Violent Crime	Murder Rate	Poverty Rate	Single Parent
AK	761	9.0	9.1	14.3	MT	178	3.0	14.9	10.8
AL	780	11.6	17.4	11.5	NC	679	11.3	14.4	11.1
AR	593	10.2	20.0	10.7	ND	82	1.7	11.2	8.4
AZ	715	8.6	15.4	12.1	NE	339	3.9	10.3	9.4
CA	1078	13.1	18.2	12.5	NH	138	2.0	9.9	9.2
CO	567	5.8	9.9	12.1	NJ	627	5.3	10.9	9.6
CT	456	6.3	8.5	10.1	NM	930	8.0	17.4	13.8
DE	686	5.0	10.2	11.4	NV	875	10.4	9.8	12.4
FL	1206	8.9	17.8	10.6	NY	1074	13.38	16.4	12.7
GA	723	11.4	13.5	13.0	OH	504	6.0	13.0	11.4
HI	261	3.8	8.0	9.1	OK	635	8.4	19.9	11.1
IA	326	2.3	10.3	9.0	OR	503	4.6	11.8	11.3
ID	282	2.9	13.1	9.5	PA	418	6.8	13.2	9.6
IL	960	11.42	13.6	11.5	RI	402	3.9	11.2	10.8
IN	489	7.5	12.2	10.8	SC	1023	10.3	18.7	12.3
KS	496	6.4	13.1	9.9	SD	208	3.4	14.2	9.4
KY	463	6.6	20.4	10.6	TN	766	10.2	19.6	11.2
LA	1062	20.3	26.4	14.9	TX	762	11.9	17.4	11.8
MA	805	3.9	10.7	10.9	UT	301	3.1	10.7	10.0
MD	998	12.7	9.7	12.0	VA	372	8.3	9.7	10.3
ME	126	1.6	10.7	10.6	VT	114	3.6	10.0	11.0
MI	792	9.8	15.4	13.0	WA	515	5.2	12.1	11.7
MN	327	3.4	11.6	9.9	WI	264	4.4	12.6	10.4
MO	744	11.3	16.1	10.9	WV	208	6.9	22.2	9.4
MS	434	13.5	24.7	14.7	WY	286	3.4	13.3	10.8
					DC	2922	78.5	26.4	22.1

LINEAR FUNCTIONS: INTERPRETING THE y -INTERCEPT AND SLOPE

Any particular formula might provide a good description or a poor one of how y relates to x . This chapter introduces the simplest type of formula—a *straight line*. For it, y is said to be a **linear function** of x .

Linear Function

The formula $y = \alpha + \beta x$ expresses observations on y as a **linear function** of observations on x . The formula has a straight-line graph with **slope** β (beta) and **y -intercept** α (alpha).

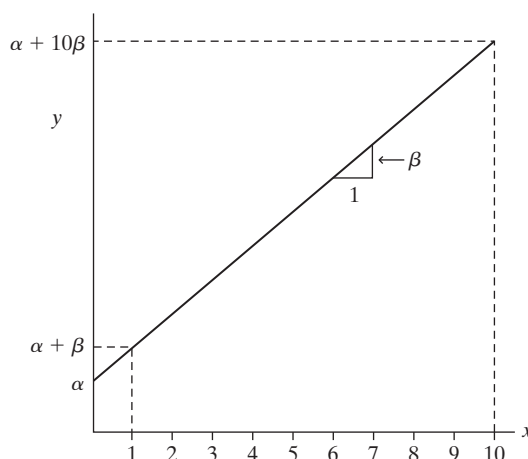
Each real number x , when substituted into the formula $y = \alpha + \beta x$, yields a distinct value for y . In a graph, the horizontal axis, the **x -axis**, lists the possible values of x . The vertical axis, the **y -axis**, lists the possible values of y . The axes intersect at the point where $x = 0$ and $y = 0$, called the *origin*.

At $x = 0$, the equation $y = \alpha + \beta x$ simplifies to $y = \alpha + \beta x = \alpha + \beta(0) = \alpha$. Thus, the constant α in this equation is the value of y when $x = 0$. Now, points on the y -axis have $x = 0$, so the line has height α at the point of its intersection with the y -axis. Because of this, α is called the **y -intercept**.

The **slope** β equals the change in y for a one-unit increase in x . That is, for two x -values that differ by 1.0 (such as $x = 0$ and $x = 1$), the y -values differ by β . Two x -values that are 10 units apart differ by 10β in their y -values. Figure 9.1 portrays the

interpretation of the y -intercept and slope. In the context of a regression analysis, α and β are called **regression coefficients**.

FIGURE 9.1: Graph of the Straight Line $y = \alpha + \beta x$. The y -intercept is α and the slope is β .



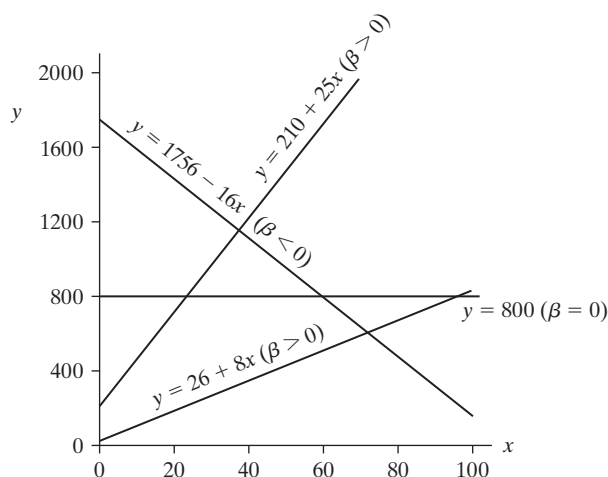
Example 9.2

Straight Lines for Predicting Violent Crime Rate For the 50 states, consider y = violent crime rate and x = poverty rate. We'll see that a straight line cannot perfectly represent the relation between them, but the line $y = 210 + 25x$ provides a type of approximation. The y -intercept of 210 represents the violent crime rate at poverty rate $x = 0$ (unfortunately, there are no such states). The slope equals 25. When the percentage with income below the poverty level increases by 1, the violent crime rate increases by about 25 crimes a year per 100,000 population.

By contrast, if instead x = percentage of the population living in urban areas, the straight line approximating the relationship is $y = 26 + 8x$. The slope of 8 is smaller than the slope of 25 when poverty rate is the explanatory variable. An increase of 1 in the percentage below the poverty level corresponds to a greater change in the violent crime rate than an increase of 1 in the percentage urban.

Figure 9.2 shows the lines relating the violent crime rate to poverty rate and to urban residence. Generally, the larger the absolute value of β , the steeper the line. When β is positive, y *increases* as x *increases*—the straight line goes upward, like these two lines. Then, large values of y occur with large values of x , and small values

FIGURE 9.2: Graphs of Lines Showing Positive Relationships ($\beta > 0$), a Negative Relationship ($\beta < 0$), and Independence ($\beta = 0$)



of y occur with small values of x . When a relationship between two variables follows a straight line with $\beta > 0$, the relationship is said to be **positive**.

When β is negative, y *decreases* as x *increases*. The straight line then goes downward, and the relationship is said to be **negative**. For instance, the equation $y = 1756 - 16x$, which has slope -16 , approximates the relationship between y = violent crime rate and x = percentage of residents who are high school graduates. For each increase of 1.0 in the percentage who are high school graduates, the violent crime rate decreases by about 16. Figure 9.2 also shows this line. ■

When $\beta = 0$, the graph is a horizontal line. The value of y is constant and does not vary as x varies. If two variables are independent, with the value of y not depending on the value of x , a straight line with $\beta = 0$ represents their relationship. The line $y = 800$ shown in Figure 9.2 is an example of a line with $\beta = 0$.

MODELS ARE SIMPLE APPROXIMATIONS FOR REALITY

As Section 7.5 (page 193) explained, a **model** is a simple approximation for the relationship between variables in the population. The linear function provides a simple model for the relationship between two quantitative variables. For a given value of x , the model $y = \alpha + \beta x$ predicts a value for y . The better these predictions tend to be, the better the model.

As we shall discuss in some detail in Chapter 10, *association does not imply causation*. For example, consider the interpretation of the slope in Example 9.2 above of “When the percentage with income below the poverty level increases by 1, the violent crime rate increases by about 25 crimes a year per 100,000 population.” This does not mean that if we had the ability to go to a state and increase the percentage of people living below the poverty level from 10% to 11%, we could expect the number of crimes to increase in the next year by 25 crimes per 100,000 people. It merely means that based on current data, if one state had a 10% poverty rate and one had an 11% poverty rate, we’d predict that the state with the higher poverty rate would have 25 more crimes per year per 100,000 people. But, as we’ll see in Section 9.3, a *sensible* model is actually a bit more complex than the one we’ve presented so far, by allowing *variability* in y -values at each value for x . That model, not merely a straight line, is what we mean by a *regression model*. Before introducing the complete model, in Section 9.3, we’ll next see how to find the best approximating straight line.

9.2 Least Squares Prediction Equation

Using sample data, we can estimate the equation for the simple straight-line model. The process treats α and β in the equation $y = \alpha + \beta x$ as parameters and estimates them.

A SCATTERPLOT PORTRAYS THE DATA

The first step of model fitting is to plot the data, to reveal whether a model with a straight-line trend makes sense. The data values (x, y) for any one subject form a point relative to the x and y axes. A plot of the n observations as n points is called a **scatterplot**.

**Example
9.3**

Scatterplot for Murder Rate and Poverty For Table 9.1, let x = poverty rate and y = murder rate. Figure 9.3 shows a scatterplot for the 51 observations. Each point portrays the values of poverty rate and murder rate for a given state. For Maryland, for instance, the poverty rate is $x = 9.7$, and the murder rate is $y = 12.7$. Its point $(x, y) = (9.7, 12.7)$ has coordinate 9.7 for the x -axis and 12.7 for the y -axis. This point is labeled MD in Figure 9.3.

FIGURE 9.3: Scatterplot for y = Murder Rate and x = Percentage of Residents below the Poverty Level, for 50 States and D.C.

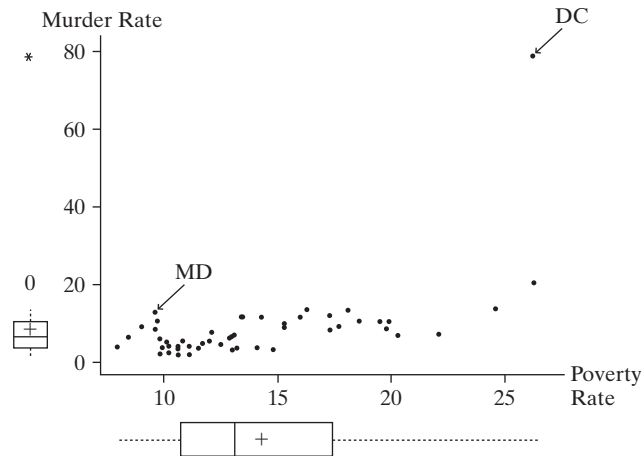
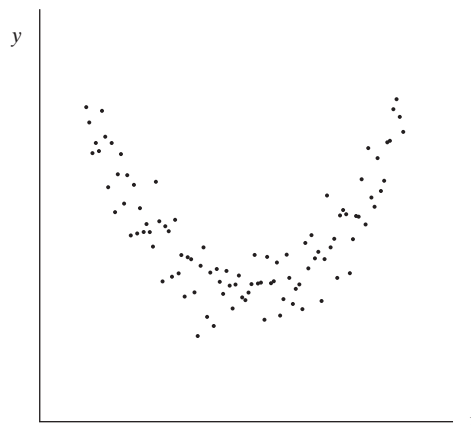


Figure 9.3 indicates that the trend of points seems to be approximated fairly well by a straight line. One point, however, is far removed from the rest. This is the point for the District of Columbia (D.C.). It had murder rate much higher than for any state. This point lies far from the overall trend. Figure 9.3 also shows box plots for these variables. They reveal that D.C. is an extreme *outlier* on murder rate. In fact, it falls 6.5 standard deviations above the mean. We shall see that outliers can have a serious impact on a regression analysis. ■

The scatterplot provides a visual check of whether a relationship is approximately linear. When the relationship seems highly nonlinear, it is not sensible to use a straight-line model. Figure 9.4 illustrates such a case. This figure shows a negative relationship over part of the range of x -values, and a positive relationship over the rest. These cancel each other out using a straight-line model. For such data, a nonlinear model presented in Section 14.5 is more appropriate.

FIGURE 9.4: A Nonlinear Relationship, for Which It Is Inappropriate to Use a Straight-Line Model



PREDICTION EQUATION

When the scatterplot suggests that the model $y = \alpha + \beta x$ may be appropriate, we use the data to estimate this line. The notation

$$\hat{y} = a + bx$$

represents a *sample* equation that estimates the linear model. In the sample equation, the y -intercept (a) estimates the y -intercept α of the model and the slope (b) estimates the slope β . The sample equation $\hat{y} = a + bx$ is called the ***prediction equation***, because it provides a prediction \hat{y} for the response variable at any value of x .

The prediction equation is the best straight line, falling closest to the points in the scatterplot, in a sense explained later in this section. The formulas for a and b in the prediction equation $\hat{y} = a + bx$ are

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} \quad \text{and} \quad a = \bar{y} - b\bar{x}.$$

If an observation has both x - and y -values above their means, or both x - and y -values below their means, then $(x - \bar{x})(y - \bar{y})$ is positive. The slope estimate b tends to be positive when most observations are like this, that is, when points with large x -values also tend to have large y -values and points with small x -values tend to have small y -values.

We shall not dwell on these formulas or even illustrate how to use them, as anyone who does any serious regression modeling uses statistical software. The appendix at the end of the text provides details. Internet applets are also available.¹

Example 9.4

Predicting Murder Rate from Poverty Rate For the 51 observations on y = murder rate and x = poverty rate in Table 9.1, SPSS software provides the results shown in Table 9.2. Murder rate has $\bar{y} = 8.7$ and $s = 10.7$, indicating that it is probably highly skewed to the right. The box plot for murder rate in Figure 9.3 shows that the extreme outlying observation for D.C. contributes to this.

The estimates of α and β are listed under the heading² *B*. The estimated y -intercept is $a = -10.14$, listed opposite (*Constant*). The estimate of the slope is $b = 1.32$, listed opposite the variable name of which it is the coefficient in the prediction equation, *POVERTY*. Therefore, the prediction equation is $\hat{y} = a + bx = -10.14 + 1.32x$.

TABLE 9.2: Part of SPSS Output for Fitting Linear Model to Observations from Crime2 Data File for 50 States and D.C. on x = Percentage in Poverty and y = Murder Rate

Variable	Mean	Std Deviation		B	Std. Error
MURDER	8.727	10.718	(Constant)	-10.1364	4.1206
POVERTY	14.259	4.584	POVERTY	1.3230	0.2754

The slope $b = 1.32$ is positive. So, the larger the poverty rate, the larger is the predicted murder rate. The value 1.32 indicates that an increase of 1 in the percentage living below the poverty rate corresponds to an increase of 1.32 in the predicted murder rate.

¹ For example, the *Fit Linear Regression Model* applet at www.artofstat.com/webapps.html.

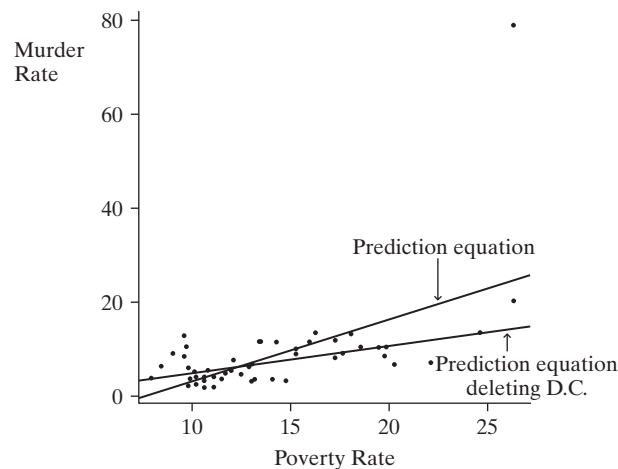
² *B* is the symbol SPSS uses to denote an estimated regression coefficient. Stata uses *Coef.* as the heading, short for coefficient, R uses *Coefficients*, and SAS uses *Parameter estimate*.

Similarly, an increase of 10 in the poverty rate corresponds to a $10(1.32) = 13.2$ -unit increase in predicted murder rate. If one state has a 12% poverty rate and another has a 22% poverty rate, for example, the predicted annual number of murders per 100,000 population is 13.2 higher in the second state than the first state. This differential of 13 murders per 100,000 population translates to 130 per million or 1300 per 10 million population. If the two states each had populations of 10 million, the one with the higher poverty rate would be predicted to have 1300 more murders per year. ■

EFFECT OF OUTLIERS ON THE PREDICTION EQUATION

Figure 9.5 plots the prediction equation from Example 9.4 over the scatterplot. The diagram shows that the observation for D.C. (the sole point in the top-right corner) is a **regression outlier**—it falls quite far from the trend that the rest of the data follow. This observation seems to have a substantial effect. The line seems to be pulled up toward it and away from the center of the general trend of points.

FIGURE 9.5: Prediction Equations Relating Murder Rate and Percentage in Poverty, with and without D.C. Observation



Let's now refit the line using the observations for the 50 states but not the one for D.C. Table 9.3 shows that the prediction equation is $\hat{y} = -0.86 + 0.58x$. Figure 9.5 also shows this line, which passes more directly through the 50 points. The slope is 0.58, compared to 1.32 when the observation for D.C. is included. The one outlying observation has the impact of more than doubling the slope!

An observation is called **influential** if removing it results in a large change in the prediction equation. Unless the sample size is large, an observation can have a strong influence on the slope if its x -value is low or high compared to the rest of the data and if it is a regression outlier.

In summary, the line for the data set including D.C. seems to distort the relationship for the 50 states. It seems wiser to use the equation based on the 50 states alone rather than to use a single equation for both the 50 states and D.C. This line for the 50 states better represents the overall trend. In reporting these results, we would note that the murder rate for D.C. falls outside this trend, being much larger than this equation predicts.

TABLE 9.3: Software Output for Fitting Linear Model to Crime2 Data File on 50 States (but Not D.C.) on x = Percentage in Poverty and y = Murder Rate

	Sum of	df	Mean	Unstandardized
	Squares		Square	Coefficients
Regression	307.342	1	307.34	B
Residual	470.406	48	9.80	(Constant) -.857
Total	777.749	49		poverty .584

	murder	predict	residual
1	9.0000	4.4599	4.5401
2	11.6000	9.3091	2.2909
3	10.2000	10.8281	-0.6281
4	8.6000	8.1406	0.4594

PREDICTION ERRORS ARE CALLED RESIDUALS

For the prediction equation $\hat{y} = -0.86 + 0.58x$ for the 50 states, a comparison of the *actual* murder rates to the *predicted* values checks the goodness of the equation. For example, Massachusetts had poverty rate $x = 10.7$ and $y = 3.9$. The predicted murder rate (\hat{y}) at $x = 10.7$ is $\hat{y} = -0.86 + 0.58x = -0.86 + 0.58(10.7) = 5.4$. The prediction error is the difference between the actual y -value of 3.9 and the predicted value of 5.4, or $y - \hat{y} = 3.9 - 5.4 = -1.5$. The prediction equation overestimates the murder rate by 1.5. Similarly, for Louisiana, $x = 26.4$ and $\hat{y} = -0.86 + 0.58(26.4) = 14.6$. The actual murder rate is $y = 20.3$, so the prediction is too low. The prediction error is $y - \hat{y} = 20.3 - 14.6 = 5.7$. The prediction errors are called **residuals**.

Residual

For an observation, the difference between an observed value and the predicted value of the response variable, $y - \hat{y}$, is called the **residual**.

Table 9.3 shows the murder rates, the predicted values, and the residuals for the first four states in the data file. A *positive* residual results when the observed value y is *larger* than the predicted value \hat{y} , so $y - \hat{y} > 0$. A *negative* residual results when the observed value is *smaller* than the predicted value. The smaller the absolute value of the residual, the better is the prediction, since the predicted value is closer to the observed value.

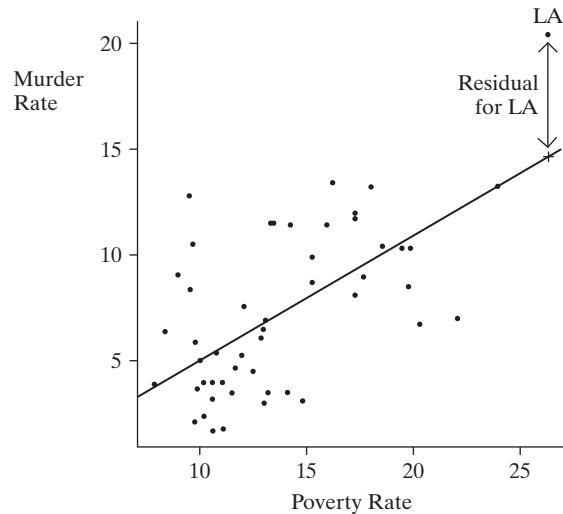
In a scatterplot, the residual for an observation is the vertical distance between its point and the prediction line. Figure 9.6 illustrates this. For example, the observation for Louisiana is the point with (x, y) coordinates $(26.4, 20.3)$. The prediction is represented by the point $(26.4, 14.6)$ on the prediction line obtained by substituting $x = 26.4$ into the prediction equation $\hat{y} = -0.86 + 0.58x$. The residual is the difference between the observed and predicted points, which is the vertical distance $y - \hat{y} = 20.3 - 14.6 = 5.7$.

PREDICTION EQUATION HAS LEAST SQUARES PROPERTY

We summarize the size of the residuals by the sum of their squared values. This quantity, denoted by SSE, is

$$\text{SSE} = \sum (y - \hat{y})^2.$$

FIGURE 9.6: Prediction Equation and Residuals. A residual is a vertical distance between a data point and the prediction line.



In other words, the residual is computed for every observation in the sample, each residual is squared, and then SSE is the sum of these squares. The symbol SSE is an abbreviation for **sum of squared errors**. This terminology refers to the residual being a measure of prediction error from using \hat{y} to predict y .

The better the prediction equation, the smaller the residuals tend to be and, hence, the smaller SSE tends to be. Any particular equation has corresponding residuals and a value of SSE. The prediction equation specified by the formulas on page 252 for the estimates a and b of α and β has the *smallest* value of SSE out of all possible linear prediction equations.

Least Squares Estimates

The **least squares estimates** a and b are the values that provide the prediction equation $\hat{y} = a + bx$ for which the residual sum of squares, $SSE = \sum(y - \hat{y})^2$, is a minimum.

The prediction line $\hat{y} = a + bx$ is called the **least squares line**, because it is the one with the smallest sum of squared residuals. If we square the residuals (such as those in Table 9.3) for the least squares line $\hat{y} = -0.86 + 0.58x$ and then sum them, we get

$$SSE = \sum(y - \hat{y})^2 = (4.54)^2 + (2.29)^2 + \cdots = 470.4.$$

This value is smaller than SSE for *any* other straight line predictor, such as $\hat{y} = -0.88 + 0.60x$. In this sense, the data fall closer to this line than to *any* other line. Most software (e.g., R, SPSS, Stata) calls SSE the **residual sum of squares**. It describes the variation of the data around the prediction line. Table 9.3 reports it in the *Sum of Squares* column, in the row labeled *Residual*.

Besides making the errors as small as possible in this summary sense, the least squares line

- Has some positive residuals and some negative residuals, but the sum (and mean) of the residuals equals 0.
- Passes through the point (\bar{x}, \bar{y}) .

The first property tells us that the too-low predictions are balanced by the too-high predictions. Just as deviations of observations from their mean \bar{y} satisfy $\sum(y - \bar{y}) = 0$,

so does the prediction equation satisfy $\sum(y - \hat{y}) = 0$. The second property tells us that the line passes through the center of the data.

9.3 The Linear Regression Model

For the linear model $y = \alpha + \beta x$, each value of x corresponds to a single value of y . Such a model is said to be **deterministic**. It is unrealistic in social science research, because we do not expect all subjects who have the same x -value to have the same y -value. Instead, the y -values *vary*.

For example, let x = number of years of education and y = annual income. The subjects having $x = 12$ years of education do not all have the same income, because income is not completely dependent upon education. Instead, a probability distribution describes annual income for individuals with $x = 12$. It is the **conditional distribution** of the y -values at $x = 12$. A separate conditional distribution applies for those with $x = 13$ years of education. Each level of education has its own conditional distribution of income. For example, the mean of the conditional distribution of income would likely be higher at higher levels of education.

A **probabilistic** model for the relationship allows for variability in y at each value of x . We now show how a linear function is the basis for a probabilistic model.

LINEAR REGRESSION FUNCTION

A probabilistic model uses $\alpha + \beta x$ to represent the *mean* of y -values, rather than y itself, as a function of x . For a given value of x , $\alpha + \beta x$ represents the mean of the conditional distribution of y for subjects having that value of x .

Expected Value of y

Let $E(y)$ denote the mean of a conditional distribution of y . The symbol E represents *expected value*.

We now use the equation

$$E(y) = \alpha + \beta x$$

to model the relationship between x and the mean of the conditional distribution of y . For y = annual income, in dollars, and x = number of years of education, suppose $E(y) = -5000 + 3000x$. For instance, those having a high school education ($x = 12$) have a mean income of $E(y) = -5000 + 3000(12) = 31,000$ dollars. The model states that the *mean* income is 31,000, but allows different subjects having $x = 12$ to have different incomes.

An equation of the form $E(y) = \alpha + \beta x$ that relates values of x to the mean of the conditional distribution of y is called a *regression function*.

Regression Function

A **regression function** is a mathematical function that describes how the mean of the response variable changes according to the value of an explanatory variable.

The function $E(y) = \alpha + \beta x$ is called a *linear regression function*, because it uses a straight line to relate the mean of y to the values of x . In practice, the *regression coefficients* α and β are unknown. Least squares provides the sample prediction equation $\hat{y} = a + bx$. At any particular value of x , $\hat{y} = a + bx$ *estimates* the mean of y for all subjects in the population having that value of x .

DESCRIBING VARIATION ABOUT THE REGRESSION LINE

The linear regression model has an additional parameter σ describing the standard deviation of each conditional distribution. That is, σ measures the variability of the y -values for all subjects having the same x -value. We refer to σ as the **conditional standard deviation**.

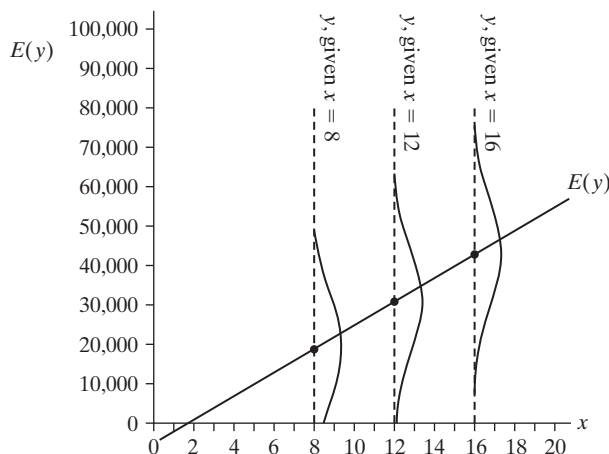
A model also assumes a particular probability distribution for the conditional distribution of y . This is needed to make inference about the parameters. For quantitative variables, the most common assumption is that the conditional distribution of y is normal at each fixed value of x , with unknown standard deviation σ .

Example
9.5

Describing How Income Varies, for Given Education Again, suppose $E(y) = -5000 + 3000x$ describes the relationship between mean annual income and number of years of education. The slope $\beta = 3000$ implies that mean income increases \$3000 for each year increase in education. Suppose also that the conditional distribution of income is normal, with $\sigma = 13,000$. According to this model, for individuals with x years of education, their incomes have a normal distribution with a mean of $E(y) = -5000 + 3000x$ and a standard deviation of 13,000.

Those having a high school education ($x = 12$) have a mean income of $E(y) = -5000 + 3000(12) = 31,000$ dollars and a standard deviation of 13,000 dollars. So, about 95% of the incomes fall within two standard deviations of the mean, that is, between $31,000 - 2(13,000) = 5000$ and $31,000 + 2(13,000) = 57,000$ dollars. Those with a college education ($x = 16$) have a mean annual income of $E(y) = -5000 + 3000(16) = 43,000$ dollars, with about 95% of the incomes falling between \$17,000 and \$69,000. Figure 9.7 pictures this regression model. ■

FIGURE 9.7: The Regression Model $E(y) = -5000 + 3000x$, with $\sigma = 13$, Relating the Mean of y = Income (in Dollars) to x = Education (in Years). The figure shows the conditional income distributions at $x = 8, 12$, and 16 years.



In Figure 9.7, each conditional distribution is normal, and each has the same standard deviation, $\sigma = 13$. In practice, the distributions would not be exactly normal, and the standard deviation need not be the same for each. *Any model never holds exactly in practice.* It is merely a simple approximation for reality. For sample data, we'll learn about ways to check whether a particular model is realistic. The most important assumption is that the regression equation is linear. The scatterplot helps us check whether this assumption is badly violated, as we'll discuss later in the chapter.

RESIDUAL MEAN SQUARE: ESTIMATING CONDITIONAL VARIATION

The ordinary linear regression model assumes that the standard deviation σ of the conditional distribution of y is identical at the various values of x . The estimate of σ uses $SSE = \sum (y - \hat{y})^2$, which measures sample variability about the least squares line. The estimate is

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}.$$

If the constant variation assumption is not valid, then s summarizes the *average* variability about the line.

**Example
9.6**

TV Watching and Grade Point Averages A survey³ of 50 college students in an introductory psychology class observed self-reports of y = high school GPA and x = weekly number of hours viewing television. The study reported $\hat{y} = 3.44 - 0.03x$. Software reports sums of squares shown in Table 9.4. This type of table is called an **ANOVA table**. Here, ANOVA is an acronym for *analysis of variability*. The residual sum of squares in using x to predict y was $SSE = 11.66$. The estimated conditional standard deviation is

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{11.66}{50-2}} = 0.49.$$

TABLE 9.4: Software Output of ANOVA Table for Sums of Squares in Fitting Regression Model to y = High School GPA and x = Weekly TV Watching

	Sum of Squares	df	Mean Square
Regression	3.63	1	3.63
Residual	11.66	48	.24
Total	15.29	49	

At any fixed value x of TV viewing, the model predicts that GPAs vary around a mean of $3.44 - 0.03x$ with a standard deviation of 0.49. At $x = 20$ hours a week, for instance, the conditional distribution of GPA is estimated to have a mean of $3.44 - 0.03(20) = 2.84$ and standard deviation of 0.49. ■

The term $(n - 2)$ in the denominator of s is the **degrees of freedom** (df) for the estimate. When a regression equation has p unknown parameters, then $df = n - p$. The equation $E(y) = \alpha + \beta x$ has two parameters (α and β), so $df = n - 2$. The table in the above example lists $SSE = 11.66$ and its $df = n - 2 = 50 - 2 = 48$. The ratio of these, $s^2 = 0.24$, is listed on the printout in the *Mean Square* column. Most software calls this the *residual mean square*. Its square root is the estimate of the conditional standard deviation of y , $s = \sqrt{0.24} = 0.49$. Among the names software calls this are *Root MSE* (Stata and SAS) for the square root of the mean square error, *Residual standard error* (R), and *Standard error of the estimate* (SPSS).

³ <https://www.iusb.edu/ugr-journal/static/2002/index.php>.

CONDITIONAL VARIATION TENDS TO BE LESS THAN MARGINAL VARIATION

From pages 43 and 105, a point estimate of the population standard deviation of a variable y is

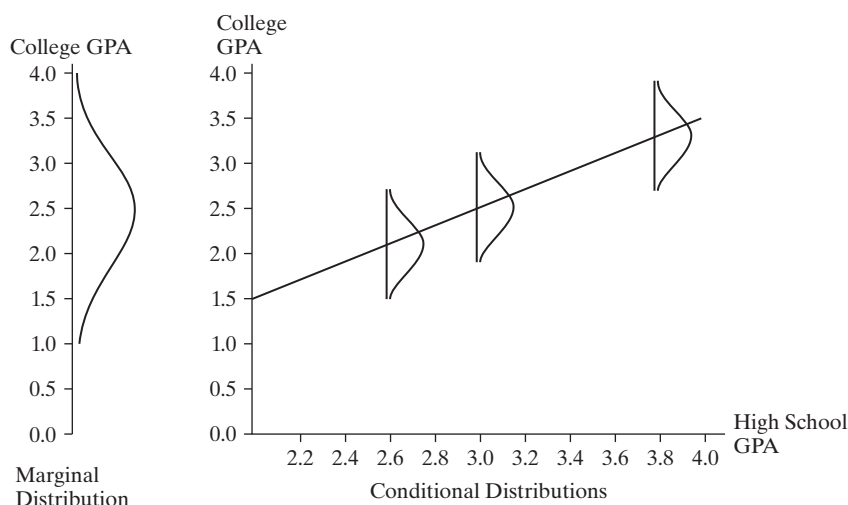
$$\sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}.$$

This is the standard deviation of the *marginal* distribution of y , because it uses only the y -values. It ignores values of x . To emphasize that this standard deviation depends on values of y alone, the remainder of the text denotes it by s_y in a sample and σ_y in a population. It differs from the standard deviation of the *conditional* distribution of y , for a fixed value of x . To reflect its conditional form, that standard deviation is sometimes denoted by $s_{y|x}$ for the sample estimate and $\sigma_{y|x}$ for the population. For simplicity, we use s and σ .

The sum of squares $\sum (y - \bar{y})^2$ in the numerator of s_y is called the **total sum of squares**. In Table 9.4 for the 50 student GPAs, it is 15.29. Thus, the marginal standard deviation of GPA is $s_y = \sqrt{15.29/(50 - 1)} = 0.56$. Example 9.6 showed that the conditional standard deviation is $s = 0.49$.

Typically, less spread in y -values occurs at a fixed value of x than totaled over all such values. We'll see that the stronger the association between x and y , the less the conditional variability tends to be relative to the marginal variability. For example, suppose the *marginal* distribution of college GPAs (y) at your school falls between 1.0 and 4.0, with $s_y = 0.60$. Suppose we could predict college GPA *perfectly* using $x =$ high school GPA, with the prediction equation $\hat{y} = 0.40 + 0.90x$. Then, $SSE = 0$, and the conditional standard deviation would be $s = 0$. In practice, perfect prediction would not happen. However, the stronger the association in terms of less prediction error, the smaller the conditional variability would be. See Figure 9.8, which portrays a marginal distribution that is much more spread out than each conditional distribution.

FIGURE 9.8: Marginal and Conditional Distributions. The marginal distribution shows the overall variability in y -values, whereas the conditional distribution shows how y varies at a fixed value of x .



9.4 Measuring Linear Association: The Correlation

The linear regression model uses a straight line to describe the relationship. For this model, this section introduces two measures of the strength of association between two quantitative variables.

THE SLOPE AND STRENGTH OF ASSOCIATION

The slope b of the prediction equation tells us the *direction* of the association. Its sign indicates whether the prediction line slopes upward or downward as x increases, that is, whether the association is positive or negative. The slope does not, however, directly tell us the strength of the association. The reason for this is that its numerical value is intrinsically linked to the units of measurement.

For example, consider the prediction equation $\hat{y} = -0.86 + 0.58x$ for $y =$ murder rate and $x =$ percentage living below the poverty level. A one-unit increase in x corresponds to a $b = 0.58$ increase in the predicted number of murders per 100,000 people. This is equivalent to a 5.8 increase in the predicted number of murders per 1,000,000 people. So, if murder rate is the number of murders per 1,000,000 population instead of per 100,000 population, the slope is 5.8 instead of 0.58. The strength of the association is the same in each case, since the variables and data are the same. Only the units of measurement for y differed. The slope b doesn't directly indicate whether the association is strong or weak, because we can make b as large or as small as we like by an appropriate choice of units.

The slope is useful for comparing effects of two predictors having the same units. For instance, the prediction equation relating murder rate to percentage living in urban areas is $3.28 + 0.06x$. A one-unit increase in the percentage living in urban areas corresponds to a 0.06 predicted increase in the murder rate, whereas a one-unit increase in the percentage below the poverty level corresponds to a 0.58 predicted increase in the murder rate. An increase of 1 in percentage below the poverty level has a much greater effect on the murder rate than an increase of 1 in percentage urban.

The measures of association we now study do not depend on the units of measurement. Like the measures of association that Chapter 8 presented for categorical data, their magnitudes indicate the strength of association.

THE CORRELATION

On page 53, we introduced the **correlation** between quantitative variables. Its value, unlike that of the slope b , does not depend on the units of measurement.

Correlation

The **correlation** between variables x and y , denoted by r , is

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\left[\sum(x - \bar{x})^2\right]\left[\sum(y - \bar{y})^2\right]}}.$$

The formulas for the correlation and for the slope (page 252) have the same numerator, relating to the covariation of x and y . The correlation is a *standardized* version of the slope. The standardization adjusts the slope b for the fact that the standard deviations of x and y depend on their units of measurement. Let s_x and s_y denote the marginal sample standard deviations of x and y ,

$$s_x = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} \quad \text{and} \quad s_y = \sqrt{\frac{\sum(y - \bar{y})^2}{n - 1}}.$$

Here is the simple connection between the slope estimate and the sample correlation:

**Correlation Is
a Standardized Slope**

The correlation relates to the slope b of the prediction equation $\hat{y} = a + bx$ by

$$r = \left(\frac{s_x}{s_y} \right) b.$$

If the sample spreads are equal ($s_x = s_y$), then $r = b$. The correlation is the value the slope would take for units such that the variables have equal standard deviations. For example, when the variables are standardized by converting their values to z -scores, both standardized variables have standard deviations of 1.0. Because of the relationship between r and b , the correlation is also called the **standardized regression coefficient** for the model $E(y) = \alpha + \beta x$. In practice, it is not necessary to standardize the variables, but we can interpret the correlation as the value the slope would equal if the variables were equally spread out.

The point estimate r of the correlation was proposed by the British statistical scientist Karl Pearson in 1896, just four years before he developed the chi-squared test of independence for contingency tables. In fact, this estimate is sometimes called the **Pearson correlation**.

**Example
9.7**

Correlation between Murder Rate and Poverty Rate For the data in Table 9.1, the prediction equation relating $y =$ murder rate to $x =$ poverty rate is $\hat{y} = -0.86 + 0.58x$. Software tells us that $s_x = 4.29$ for poverty rate, $s_y = 3.98$ for murder rate, and the correlation $r = 0.63$. In fact,

$$r = \left(\frac{s_x}{s_y} \right) b = \left(\frac{4.29}{3.98} \right) (0.58) = 0.63.$$

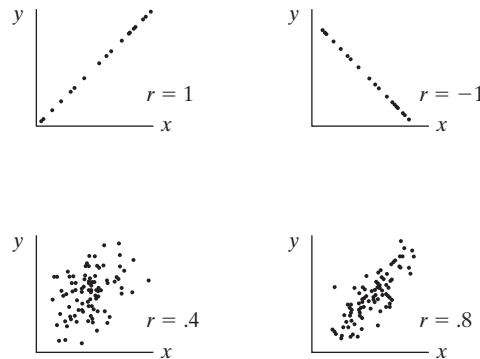
We will interpret this value after studying the properties of the correlation. ■

PROPERTIES OF THE CORRELATION

- The correlation is valid only when a straight-line model is sensible for the relationship between x and y . Since r is proportional to the slope of a linear prediction equation, it measures the *strength of the linear association*.
- $-1 \leq r \leq 1$. The correlation, unlike the slope b , must fall between -1 and $+1$. We shall see why later in the section.
- r has the same sign as the slope b . This holds because their formulas have the same numerator, relating to covariation of x and y , and positive denominators. Thus, $r > 0$ when the variables are positively related, and $r < 0$ when the variables are negatively related.
- $r = 0$ for those lines having $b = 0$. When $r = 0$, there is not a linear increasing or linear decreasing trend in the relationship.
- $r = \pm 1$ when all the sample points fall exactly on the prediction line. These correspond to *perfect* positive and negative linear associations. There is then no prediction error when we use $\hat{y} = a + bx$ to predict y .

- The larger the absolute value of r , the stronger the linear association. Variables with a correlation of -0.80 are more strongly linearly associated than variables with a correlation of 0.40 . Figure 9.9 shows scatterplots having various values for r .

FIGURE 9.9: Scatterplots for Different Correlation Values



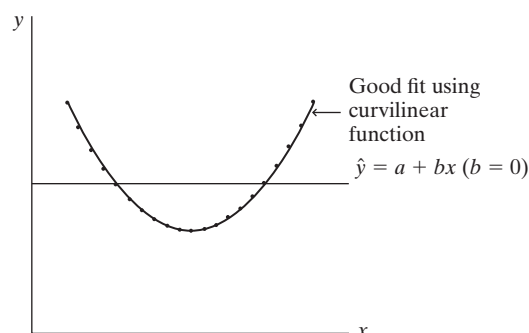
- The correlation, unlike the slope b , treats x and y symmetrically. The prediction equation using y to predict x has the same correlation as the one using x to predict y .
- The value of r does not depend on the variables' units.

For example, if y is the number of murders per 1,000,000 population instead of per 100,000 population, we obtain the same value of $r = 0.63$. Also, when murder rate predicts poverty rate, the correlation is the same as when poverty rate predicts murder rate, $r = 0.63$ in both cases.

The correlation is useful for comparing associations for variables having different units. Another potential predictor for murder rate is the mean number of years of education completed by adult residents in the state. Poverty rate and education have different units, so a one-unit change in poverty rate is not comparable to a one-unit change in education. Their slopes from the separate prediction equations are not comparable. The correlations are comparable. Suppose the correlation of murder rate with education is -0.30 . Since the correlation of murder rate with poverty rate is 0.63 , and since $0.63 > |-0.30|$, murder rate is more strongly associated with poverty rate than with education.

We emphasize that the correlation describes *linear* relationships. For curvilinear relationships, the best-fitting prediction line may be completely or nearly horizontal, and $r = 0$ when $b = 0$. See Figure 9.10. A low absolute value for r does not then imply that the variables are unassociated, but merely that the association is not linear.

FIGURE 9.10: Scatterplot for Which $r = 0$, Even Though There Is a Strong Curvilinear Relationship



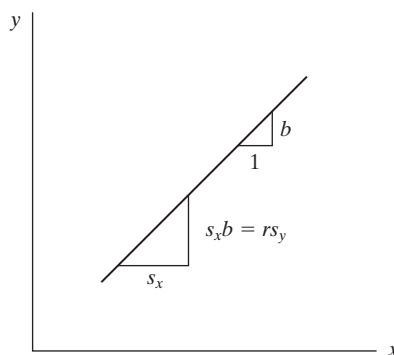
CORRELATION IMPLIES REGRESSION TOWARD THE MEAN

Another interpretation of the correlation relates to its standardized slope property. We can rewrite the equality

$$r = \left(\frac{s_x}{s_y} \right) b \quad \text{as} \quad s_x b = r s_y.$$

Now, the slope b is the change in \hat{y} for a one-unit increase in x . An increase in x of s_x units has a predicted change of $s_x b$ units. (For instance, if $s_x = 10$, an increase of 10 units in x corresponds to a change in \hat{y} of $10b$.) See Figure 9.11. Since $s_x b = r s_y$, an increase of s_x in x corresponds to a predicted change of r standard deviations in the y -values.

FIGURE 9.11: An Increase of s_x Units in x Corresponds to a Predicted Change of $r s_y$ Units in y



For example, let's start at the point (\bar{x}, \bar{y}) through which the prediction equation passes and consider the impact of x moving above \bar{x} by a standard deviation. Suppose that $r = 0.5$. An increase of s_x in x , from \bar{x} to $(\bar{x} + s_x)$, corresponds to a predicted increase of $0.5s_y$ in y , from \bar{y} to $(\bar{y} + 0.5s_y)$. We predict that y is closer to the mean, in standard deviation units. This is called **regression toward the mean**. The larger the absolute value of r , the stronger the association, in the sense that a standard deviation change in x corresponds to a greater proportion of a standard deviation change in y .

Example
9.8

Child's Height Regresses toward the Mean The British scientist Sir Francis Galton discovered the basic ideas of regression and correlation in the 1880s. After multiplying each female height by 1.08 to account for gender differences, he noted that the correlation between x = parent height (the average of father's and mother's height) and y = child's height is about 0.5. From the property just discussed, a standard deviation change in parent height corresponds to half a standard deviation change in child's height.

For parents of average height, the child's height is predicted to be average. If, on the other hand, parent height is a standard deviation above average, the child is predicted to be half a standard deviation above average. If parent height is two standard deviations below average, the child is predicted to be one standard deviation below average.

Since r is less than 1, a y -value is predicted to be fewer standard deviations from its mean than x is from its mean. Tall parents tend to have tall children, but on the average not quite so tall. For instance, if you consider all fathers with height 7 feet, perhaps their sons average 6 feet 5 inches—taller than average, but not so extremely tall; if you consider all fathers with height 5 feet, perhaps their sons average 5 feet

5 inches—shorter than average, but not so extremely short. In each case, Galton pointed out the *regression toward the mean*. This is the origin of the name for regression analysis. ■

r-SQUARED: PROPORTIONAL REDUCTION IN PREDICTION ERROR

A related measure of association summarizes how well x can predict y . If we can predict y much better by substituting x -values into the prediction equation $\hat{y} = a + bx$ than without knowing the x -values, the variables are judged to be strongly associated. This measure of association has four elements:

- A rule for predicting y without using x . We refer to this as Rule 1.
- A rule for predicting y using information on x . We refer to this as Rule 2.
- A summary measure of prediction error for each rule, E_1 for errors by rule 1 and E_2 for errors by rule 2.
- The difference in the amount of error with the two rules is $E_1 - E_2$. Converting this reduction in error to a proportion provides the definition

$$\text{Proportional reduction in error} = \frac{E_1 - E_2}{E_1}.$$

Rule 1 (Predicting y without using x): The best predictor is \bar{y} , the sample mean.

Rule 2 (Predicting y using x): When the relationship between x and y is linear, the prediction equation $\hat{y} = a + bx$ provides the best predictor of y .

Prediction Errors: The prediction error for each subject is the difference between the observed and predicted values of y . The prediction error using rule 1 is $y - \bar{y}$, and the prediction error using rule 2 is $y - \hat{y}$, the residual. For each predictor, some prediction errors are positive, some are negative, and the sum of the errors equals 0. We summarize the prediction errors by their sum of squared values,

$$E = \sum (\text{observed } y\text{-value} - \text{predicted } y\text{-value})^2.$$

For rule 1, the predicted values all equal \bar{y} . The total prediction error is

$$E_1 = \sum (y - \bar{y})^2.$$

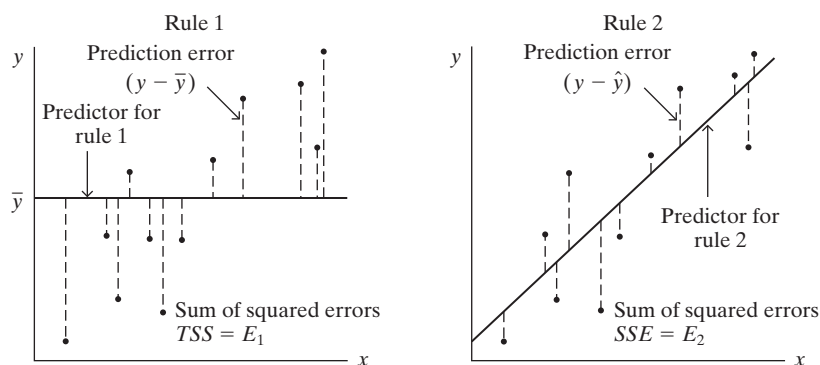
This is the *total sum of squares* of the y -values about their mean. We denote this by TSS. For rule 2 (predicting using the \hat{y} -values), the total prediction error is

$$E_2 = \sum (y - \hat{y})^2.$$

We have denoted this by SSE, called the *sum of squared errors* or the *residual sum of squares*.

When x and y have a strong linear association, the prediction equation provides predictions (\hat{y}) that are much better than \bar{y} , in the sense that the sum of squared prediction errors is substantially less. Figure 9.12 shows graphical representations of the two predictors and their prediction errors. For rule 1, the same prediction (\bar{y}) applies for the value of y , regardless of the value of x . For rule 2, the prediction changes as x changes, and the prediction errors tend to be smaller.

FIGURE 9.12: Graphical Representation of Rule 1 and Total Sum of Squares $E_1 = \text{TSS} = \sum(y - \bar{y})^2$, Rule 2 and Residual Sum of Squares $E_2 = \text{SSE} = \sum(y - \hat{y})^2$



Definition of Measure: The proportional reduction in error from using the linear prediction equation instead of \bar{y} to predict y is

$$r^2 = \frac{E_1 - E_2}{E_1} = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = \frac{\sum(y - \bar{y})^2 - \sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}.$$

It is called ***r-squared***, or sometimes the ***coefficient of determination***. The notation r^2 is used for this measure because, in fact, the proportional reduction in error equals the square of the correlation r .

Example 9.9

r^2 for Murder Rate and Poverty Rate The correlation between poverty rate and murder rate for the 50 states is $r = 0.629$. Therefore, $r^2 = (0.629)^2 = 0.395$. For predicting murder rate, the linear prediction equation $\hat{y} = -0.86 + 0.58x$ has 39.5% less error than \bar{y} .

Software for regression routinely provides tables that contain the sums of squares that compose r^2 . For example, part of Table 9.3 contained the ANOVA table

	Sum of Squares
Regression	307.342
Residual	470.406
Total	777.749

The sum of squared errors using the prediction equation is $\text{SSE} = \sum(y - \hat{y})^2 = 470.4$, and the total sum of squares is $\text{TSS} = \sum(y - \bar{y})^2 = 777.7$. Thus,

$$r^2 = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = \frac{777.7 - 470.4}{777.7} = \frac{307.3}{777.7} = 0.395.$$

In practice, it is unnecessary to perform this computation, since software reports r or r^2 or both. ■

PROPERTIES OF r -SQUARED

The properties of r^2 follow directly from those of the correlation r or from its definition in terms of the sums of squares.

- Since $-1 \leq r \leq 1$, r^2 falls between 0 and 1.
- The minimum possible value for SSE is 0, in which case $r^2 = \text{TSS}/\text{TSS} = 1$. For $\text{SSE} = 0$, all sample points must fall exactly on the prediction line. In that case, there is no error using x to predict y with the prediction equation. This condition corresponds to $r = \pm 1$.

- When the least squares slope $b = 0$, the y -intercept a equals \bar{y} (because $a = \bar{y} - b\bar{x}$, which equals \bar{y} when $b = 0$). Then, $\hat{y} = \bar{y}$ for all x . The two prediction rules are then identical, so $SSE = TSS$ and $r^2 = 0$.
- Like the correlation, r^2 measures the strength of *linear* association. The closer r^2 is to 1, the stronger the linear association, in the sense that the more effective the least squares line $\hat{y} = a + bx$ is compared to \bar{y} in predicting y .
- r^2 does not depend on the units of measurement, and it takes the same value when x predicts y as when y predicts x .

SUMS OF SQUARES DESCRIBE CONDITIONAL AND MARGINAL VARIABILITY

To summarize, the correlation r falls between -1 and $+1$. It indicates the direction of the association, positive or negative, through its sign. It is a standardized slope, equaling the slope when x and y are equally spread out. A one standard deviation change in x corresponds to a predicted change of r standard deviations in y . The square of the correlation has a proportional reduction in error interpretation related to predicting y using $\hat{y} = a + bx$ rather than \bar{y} .

The total sum of squares, $TSS = \sum(y - \bar{y})^2$, summarizes the *variability* of the observations on y , since this quantity divided by $n - 1$ is the sample variance s_y^2 of the y -values. Similarly, $SSE = \sum(y - \hat{y})^2$ summarizes the variability around the prediction equation, which refers to variability for the conditional distributions. For example, when $r^2 = 0.39$, the variability in y using x to make the predictions is 39% less than the overall variability of the y -values. Thus, the r^2 result is often expressed as “the poverty rate explains 39% of the variability in murder rate” or “39% of the variance in murder rate is explained by its linear relationship with the poverty rate.” Roughly speaking, the variance of the conditional distribution of murder rate for a given poverty rate is 39% smaller than the variance of the marginal distribution of murder rate.

This interpretation has the weakness, however, that variability is summarized by the *variance*. Many statisticians find r^2 to be less useful than r because, being based on sums of squares, it uses the square of the original scale of measurement. It’s easier to interpret the original scale than a squared scale. This is also the advantage of the standard deviation over the variance.

9.5 Inferences for the Slope and Correlation

We have seen that a linear regression model can represent the *form* of a relationship between two quantitative variables. We use the correlation and its square to describe the *strength* of the association. These parts of a regression analysis are descriptive. We now present inferential methods for the regression model.

A test of whether the two quantitative variables are statistically independent has the same purpose as the chi-squared test for categorical variables. A confidence interval for the slope of the regression equation or the correlation tells us about the size of the effect. These inferences enable us to judge whether the variables are associated and to estimate the direction and strength of the association.

ASSUMPTIONS FOR STATISTICAL INFERENCE

Statistical inferences for regression make the following assumptions:

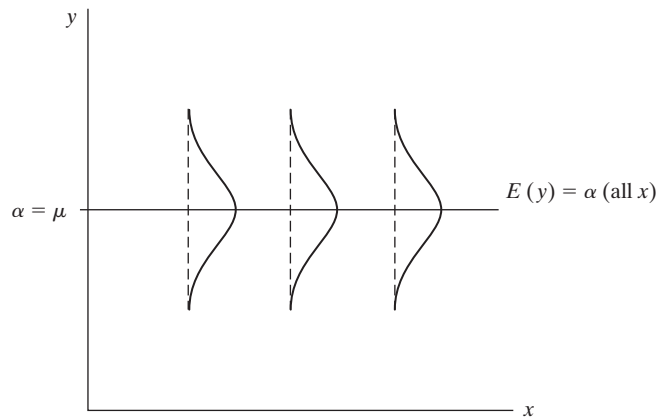
- Randomization, such as a simple random sample in a survey.
- The mean of y is related to x by the linear equation $E(y) = \alpha + \beta x$.
- The conditional standard deviation σ is identical at each x -value.
- The conditional distribution of y at each value of x is normal.

The assumption about a common σ is one under which the least squares estimates are the best possible estimates of the regression coefficients.⁴ The assumption about normality assures that the test statistic for a test of independence has a t sampling distribution. In practice, none of these assumptions is ever satisfied exactly. In the final section of the chapter, we'll see that the important assumptions are the first two.

TEST OF INDEPENDENCE USING SLOPE OR CORRELATION

Under the above assumptions, suppose the population mean of y is identical at each x -value. In other words, the normal conditional distribution of y is the same at each x -value. Then, the two quantitative variables are statistically independent. For the linear regression function $E(y) = \alpha + \beta x$, this means that the slope $\beta = 0$ (see Figure 9.13). The null hypothesis that the variables are statistically independent is $H_0: \beta = 0$.

FIGURE 9.13: x and y Are Statistically Independent when the Slope $\beta = 0$ in the Regression Model $E(y) = \alpha + \beta x$ with Normal Conditional Distributions Having Identical Standard Deviations



We can test independence against $H_a: \beta \neq 0$, or a one-sided alternative, $H_a: \beta > 0$ or $H_a: \beta < 0$, to predict the direction of the association. The test statistic is

$$t = \frac{b}{se},$$

where se is the standard error of the sample slope b . The form of the test statistic is the usual one for a t or z test. We take the estimate b of the parameter β , subtract the null hypothesis value ($\beta = 0$), and divide by the standard error of the estimate b . Under the assumptions, this test statistic has the t sampling distribution with $df = n - 2$. The P -value for $H_a: \beta \neq 0$ is the two-tail probability from the t distribution.

The formula for the standard error of b is

$$se = \frac{s}{\sqrt{\sum (x - \bar{x})^2}}, \quad \text{where} \quad s = \sqrt{\frac{SSE}{n - 2}}.$$

⁴ Under the assumptions of normality with common σ , least squares estimates are special cases of *maximum likelihood* estimates, introduced in Section 5.5.

This depends on the point estimate s of the standard deviation of the conditional distributions of y . The degrees of freedom for the t test are the same as the df for s . The smaller s is, the more precisely b estimates β . A small s occurs when the data points show little variability about the prediction equation. Also, the standard error of b is inversely related to $\sum(x - \bar{x})^2$, the sum of squares of the observed x -values about their mean. This sum increases, and hence b estimates β more precisely, as the sample size n increases. (The se also decreases when the x -values are more highly spread out, but the researcher usually has no control over this except in designed experiments.)

The correlation $r = 0$ in the same situations in which the slope $b = 0$. Let ρ (Greek letter rho) denote the correlation value in the population. Then, $\rho = 0$ precisely when $\beta = 0$. In fact, a test of $H_0: \rho = 0$ using the sample value r is equivalent to the t test of $H_0: \beta = 0$ using the sample value b . The test statistic for $H_0: \rho = 0$ is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}.$$

This has the same value as the test statistic $t = b/se$. We can use either statistic to test H_0 : independence, since each has the t distribution with $df = n - 2$ and yields the same P -value.

Example
9.10

Regression for Selling Price of Homes What affects the selling price of a house? Table 9.5 shows observations on recent home sales in Gainesville, Florida. This table shows data for eight homes. The entire file for 100 home sales is the **Houses** data file at the text website. Variables listed are selling price (in dollars), size of house (in square feet), annual property taxes (in dollars), number of bedrooms, number of bathrooms, and whether the house is newly built.

TABLE 9.5: Selling Prices and Related Factors for a Sample of Home Sales in Gainesville, Florida (Houses Data File)						
Home	Selling Price	Size	Taxes	Bedrooms	Bathrooms	New
1	279,900	2048	3104	4	2	No
2	146,500	912	1173	2	1	No
3	237,700	1654	3076	4	2	No
4	200,000	2068	1608	3	2	No
5	159,900	1477	1454	3	3	No
6	499,900	3153	2997	3	2	Yes
7	265,500	1355	4054	3	2	No
8	289,900	2075	3002	3	2	Yes

Note: The complete Houses data file for 100 homes is at the text website.

For a set of variables, software can report the correlation for each pair in a **correlation matrix**. This matrix is a square table listing the variables as the rows and again as the columns. Table 9.6 shows the way software reports the correlation matrix for the variables selling price, size, taxes, and number of bedrooms. The correlation between each pair of variables appears twice. For instance, the correlation of 0.834 between selling price and size of house occurs both in the row for *price* and column for *size* and in the row for *size* and column for *price*. The correlations on the diagonal running from the upper left-hand corner to the lower right-hand corner of a correlation matrix all equal 1.0. This merely indicates that the correlation between a variable and itself is 1.0. For instance, if we know the value of y , then we can predict the value of y perfectly.

TABLE 9.6: Correlation Matrix for House Selling Price Data from Houses Data File

	Correlations			
	price	size	taxes	bedrooms
price	1.00000	0.83378	0.84198	0.39396
size	0.83378	1.00000	0.81880	0.54478
taxes	0.84198	0.81880	1.00000	0.47393
bedrooms	0.39396	0.54478	0.47393	1.00000

For now, we use only the data on y = selling price and x = size of house. Since these 100 observations come from one city alone, we cannot use them to make inferences about the relationship between x and y in general. We treat them as a random sample of a conceptual population of home sales in this market in order to analyze how these variables seem to be related.

Figure 9.14 shows a scatterplot, which displays a strong positive trend. The model $E(y) = \alpha + \beta x$ seems appropriate. Some of the points at high levels of size may be outliers, however, and one point falls quite far below the overall trend. We discuss this abnormality in Section 14.4, which introduces an alternative model that does not assume constant variability around the regression line.

FIGURE 9.14: Scatterplot and Prediction Equation for y = Selling Price (in Dollars) and x = Size of House (in Square Feet)

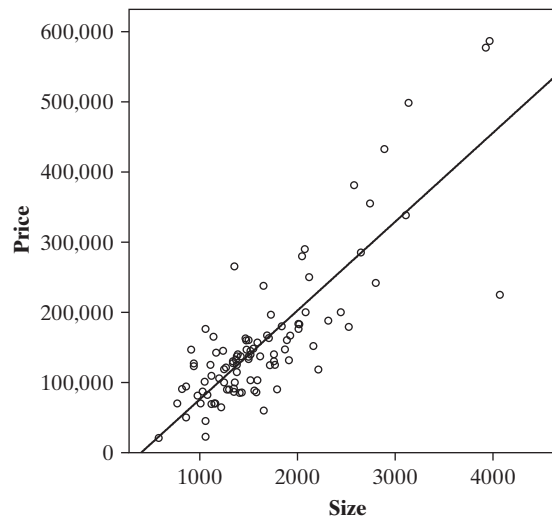


Table 9.7 shows some software output (Stata) for a regression analysis. The prediction equation is $\hat{y} = -50,926.2 + 126.6x$. The predicted selling price increases by $b = 126.6$ dollars for an increase in size of a square foot. Figure 9.14 also superimposes the prediction equation over the scatterplot.

Table 9.7 reports that the standard error of the slope estimate is $se = 8.47$. This value estimates the variability in sample slope values that would result from repeatedly selecting random samples of 100 house sales in Gainesville and calculating prediction equations. For testing independence, $H_0: \beta = 0$, the test statistic is

$$t = \frac{b}{se} = \frac{126.6}{8.47} = 14.95,$$

shown in Table 9.7. Since $n = 100$, its degrees of freedom are $df = n - 2 = 98$. This is an extremely large test statistic. The P -value, listed in Table 9.7 under the heading

TABLE 9.7: Stata Output (Edited) for Regression Analysis of $y = \text{Selling Price}$ and $x = \text{Size of House}$ from `Houses` Data File

Source	SS	df	MS	Number of obs =	100
Model	7.0573e+11	1	7.0573e+11	F(1, 98)	= 223.52
Residual	3.0942e+11	98	3.1574e+09	Prob > F	= 0.0000
Total	1.0151e+12	99	1.0254e+10	R-squared	= 0.6952
				Adj R-squared	= 0.6921
				Root MSE	= 56190
price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
size	126.5941	8.467517	14.95	0.000	109.7906 143.3976
_cons	-50926.25	14896.37	-3.42	0.001	-80487.62 -21364.89

$P > |t|$, is 0.000 to three decimal places. This refers to the two-sided alternative $H_a: \beta \neq 0$. It is the two-tailed probability of a t statistic at least as large in absolute value as the absolute value of the observed t , $|t| = 14.95$, presuming H_0 is true.

We get the same result if we conduct the test using the correlation. The correlation of $r = 0.834$ for the house selling price data has

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.834}{\sqrt{(1-0.695)/98}} = 14.95.$$

Table 9.8 shows some R output for the same analysis. The two-sided P -value, listed under the heading $Pr(> |t|)$, is 0 to many decimal places.

TABLE 9.8: R Output for Regression Analysis of $y = \text{Selling Price}$ and $x = \text{Size of House}$ from `Houses` Data File

```
> fit <- lm(price ~ size)
> summary(fit)

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -50926.255  14896.373  -3.419 0.000918
size         126.594    8.468   14.951 < 2e-16
---
Residual standard error: 56190 on 98 degrees of freedom
Multiple R-squared:  0.6952,    Adjusted R-squared:  0.6921

> cor.test(price,size)

t = 14.951, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:  0.7621910 0.8852286
sample estimates:  cor  0.8337848
```

Both the Stata and R outputs also contain a standard error and t test for the y -intercept. We won't use this information, since rarely is there any reason to test the hypothesis that a y -intercept equals 0. For this example, the y -intercept does not have any interpretation, since houses of size $x = 0$ do not exist.

In summary, the evidence is extremely strong that size of house has a positive effect on selling price. On the average, selling price increases as size increases. This

is no surprise. Indeed, we would be shocked if these variables were independent. For these data, estimating the size of the effect is more relevant than testing whether it exists. ■

CONFIDENCE INTERVAL FOR THE SLOPE AND CORRELATION

A small P -value for $H_0: \beta = 0$ suggests that the regression line has a nonzero slope. We should be more concerned with the size of the slope β than in knowing merely that it is not 0. A confidence interval for β has the formula

$$b \pm t(se).$$

The t -score is the value, with $df = n - 2$, for the desired confidence level. The form of the interval is similar to the confidence interval for a mean (Section 5.3), namely, take the estimate of the parameter and add and subtract a t multiple of the standard error. The se is the same as se in the test about β .

Constructing a confidence interval for the correlation ρ is more complicated than for the slope β . The reason is that the sampling distribution of r is not symmetric except when $\rho = 0$. The lack of symmetry is caused by the restricted range $[-1, 1]$ for r values. If ρ is close to 1.0, for instance, the sample r cannot fall much above ρ , but it can fall well below ρ . The sampling distribution of r is then skewed to the left. Exercise 9.64 shows how to construct confidence intervals for correlations. This is available with software.

Example 9.11

Estimating the Slope and Correlation for House Selling Prices For the data on x = size of house and y = selling price, $b = 126.6$ and $se = 8.47$. The parameter β refers to the change in the mean selling price (in dollars) for each 1-square-foot increase in size. For a 95% confidence interval, we use the $t_{.025}$ value for $df = n - 2 = 98$, which is $t_{.025} = 1.984$. The interval is

$$\begin{aligned} b \pm t_{.025}(se) &= 126.6 \pm 1.984(8.47) \\ &= 126.6 \pm 16.8, \quad \text{or} \quad (110, 143). \end{aligned}$$

We can be 95% confident that β falls between 110 and 143. The mean selling price increases by between \$110 and \$143 for a 1-square-foot increase in house size. ■

In practice, we make inferences about the change in $E(y)$ for an increase in x that is a relevant portion of the actual range of x -values. If a one-unit increase in x is too small or too large in practical terms, the confidence interval for β can be adjusted to refer to a different change in x . For Table 9.5, x = size of house has $\bar{x} = 1629$ and $s_x = 669$. A change of 1 square foot in size is small. Let's estimate the effect of a 100-square-foot increase in size. The change in the mean of y is 100β . The 95% confidence interval for β is (110, 143), so the 95% confidence interval for 100β has endpoints $100(110) = 11,100$ and $100(143) = 14,300$. We infer that the mean selling price increases by at least \$11,100 and at most \$14,300 for a 100-square-foot increase in house size. For example, assuming that the linear regression model is valid, we conclude that the mean is between \$11,100 and \$14,300 higher for houses of 1700 square feet than for houses of 1600 square feet.

For the house selling price data, we found that the correlation between selling price and size is 0.834. The R output in Table 9.8 tells us that a 95% confidence interval for the population correlation is (0.762, 0.885).

SUMS OF SQUARES IN SOFTWARE OUTPUT

How do we interpret the sums of squares (SS) output in tables such as Table 9.7? In that table, the residual sum of squares ($SSE = 3.0942 \times 10^{11}$) is a huge number because the y -values are very large and their deviations are squared. The estimated conditional standard deviation of y is

$$s = \sqrt{SSE/(n-2)} = 56,190,$$

labeled as *Root MSE* by Stata and *Residual standard error* by R. The sum of squares table also reports the total sum of squares, $TSS = \sum(y - \bar{y})^2 = 1.0151 \times 10^{12}$. From this value and SSE,

$$r^2 = \frac{TSS - SSE}{TSS} = 0.695.$$

This is the proportional reduction in error in predicting the selling price using the linear prediction equation instead of the sample mean selling price. A strong association exists between these variables.

The total sum of squares TSS partitions into two parts, the sum of squared errors, $SSE = 3.0942 \times 10^{11}$, and the difference $TSS - SSE = 7.0573 \times 10^{11}$. This difference is the numerator of the r^2 measure. Software calls this the **regression sum of squares** (e.g., SPSS) or the **model sum of squares** (e.g., Stata, SAS). It represents the amount of the total variation TSS in y that is explained by x in using the least squares line. The ratio of this sum of squares to TSS equals r^2 .

Tables of sums of squares have an associated list of degrees of freedom values. The df for the total sum of squares $TSS = \sum(y - \bar{y})^2$ is $n - 1 = 99$, since TSS refers to variability in the *marginal* distribution of y , which has sample variance $s_y^2 = TSS/(n - 1)$. The degrees of freedom for SSE are $n - 2 = 98$, since it refers to variability in the *conditional* distribution of y , which has variance estimate $s^2 = SSE/(n - 2)$ for a model having two parameters. The regression (model) sum of squares has df equal to the number of explanatory variables in the regression model, in this case 1. The sum of df for the regression sum of squares and df for the residual sum of squared errors is $df = n - 1$ for the total sum of squares, in this case $1 + 98 = 99$.

9.6 Model Assumptions and Violations

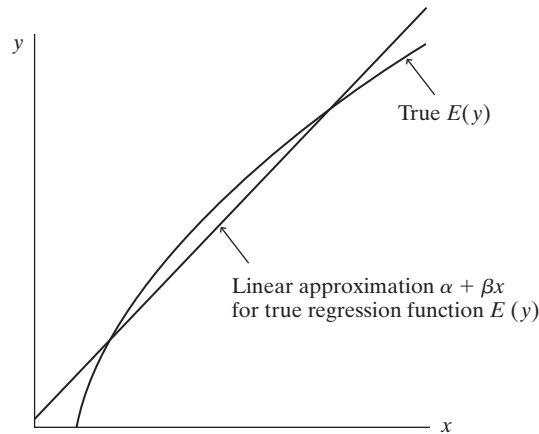
We end this chapter by reconsidering the assumptions underlying linear regression analysis. We discuss the effects of violating these assumptions and the effects of *influential* observations. Finally, we show an alternate way to express the model.

WHICH ASSUMPTIONS ARE IMPORTANT?

The linear regression model assumes that the relationship between x and the mean of y follows a straight line. The actual form is unknown. It is almost certainly not *exactly* linear. Nevertheless, a linear function often provides a decent approximation for the actual form. Figure 9.15 illustrates a straight line falling close to an actual curvilinear relationship.

The inferences introduced in the previous section are appropriate for detecting positive or negative linear associations. Suppose that instead the true relationship were U-shaped, such as in Figure 9.4. Then, the variables would be statistically dependent, since the mean of y would change as the value of x changes. The t test of $H_0: \beta = 0$ might not detect it, though, because the slope b of the least squares line would be close to 0. In other words, a small P -value would probably not occur even though

FIGURE 9.15: A Linear Regression Equation as an Approximation for a Nonlinear Relationship



an association exists. In summary, $\beta = 0$ need not correspond to independence if the assumption of a linear regression model is violated. For this reason, you should always construct a scatterplot to check this fundamental assumption.

The least squares line and r and r^2 are valid descriptive statistics no matter what the shape of the conditional distribution of y -values for each x -value. However, the statistical inferences in Section 9.5 also assume that the conditional distributions of y are (1) normal, with (2) identical standard deviation σ for each x -value. These assumptions are also not *exactly* satisfied in practice. For large samples, the normality assumption is relatively unimportant, because an extended Central Limit Theorem implies that sample slopes and correlations have approximately normal sampling distributions. If the assumption about common σ is violated, other estimates may be more efficient than least squares (i.e., having smaller *se* values), but ordinary inference methods are still approximately valid.

The random sample and straight-line assumptions are very important. If the true relationship deviates greatly from a straight line, for instance, it does not make sense to use a slope or a correlation to describe it. Chapter 14 discusses ways of checking the assumptions and modifying the analysis, if necessary.

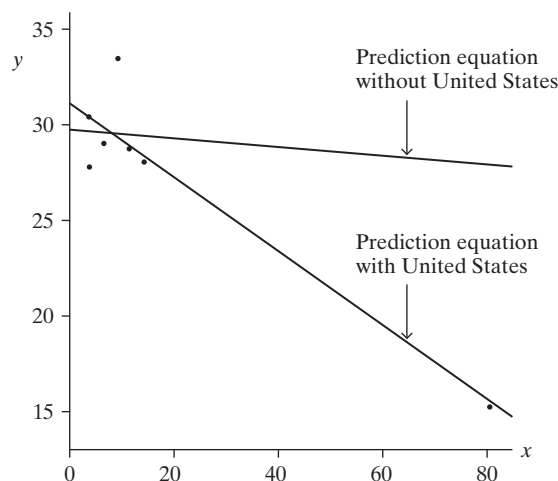
INFLUENTIAL OBSERVATIONS

The least squares method has a long history and is the standard way to fit prediction equations to data. A disadvantage of least squares, however, is that individual observations can unduly influence the results. A single observation can have a large effect if it is a *regression outlier*—having x -value relatively large or relatively small and falling quite far from the trend that the rest of the data follow.

Figure 9.16 illustrates this. The figure plots observations for several African and Asian nations on y = crude birth rate (number of births per 1000 population size) and x = number of televisions per 100 people. We added to the figure an observation on these variables for the United States, which is the outlier that is much lower than the other countries in birth rate but much higher on number of televisions. Figure 9.16 shows the prediction equations both without and with the U.S. observation. The prediction equation changes from $\hat{y} = 29.8 - 0.024x$ to $\hat{y} = 31.2 - 0.195x$. Adding only a single point to the data set causes the prediction line to tilt dramatically downward.

When a scatterplot shows a severe regression outlier, you should investigate the reasons for it. An observation may have been incorrectly recorded. If the observation is correct, perhaps that observation is fundamentally different from the others in some way, such as the U.S. observation in Figure 9.16. It may suggest an additional

FIGURE 9.16: Prediction Equations for $y = \text{Birth Rate}$ and $x = \text{Television Ownership}$, with and without Observation for the United States



predictor for the model, using methods of Chapter 11. It is often worthwhile to refit the model without one or two extreme regression outliers to see if those observations have a large effect on the fit. We did this following Example 9.4 (page 252) with the D.C. observation for the murder rates. The slope of the prediction equation relating murder rate to poverty rate more than doubled when we included the observation for D.C.

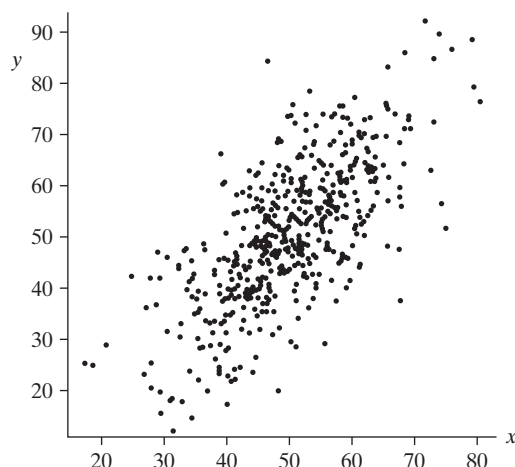
Observations that have a large influence on the model parameter estimates can also have a large impact on the correlation. For instance, for the data in Figure 9.16, the correlation is -0.935 when the United States is included and -0.051 when it is deleted from the data set. One point can make quite a difference, especially when the sample size is small.

FACTORS INFLUENCING THE CORRELATION

Besides being influenced by outliers, the correlation depends on the range of x -values sampled. When a sample has a much narrower range of variation in x than the population, the sample correlation tends to underestimate drastically (in absolute value) the population correlation.

Figure 9.17 shows a scatterplot of 500 points that has a correlation of $r = 0.71$. Suppose, instead, we had only sampled the middle half of the points, roughly between

FIGURE 9.17: The Correlation Is Affected by the Range of x -Values. The correlation decreases from 0.71 to 0.33 using only points with x between 43 and 57.



x -values of 43 and 57. Then the correlation equals only $r = 0.33$, considerably lower. For the relation between housing price and size of house, portrayed in Figure 9.14, $r = 0.834$. If we sampled only those sales in which house size is between 1300 and 2000 square feet, which include 44 of the 100 observations, r decreases to 0.254.

The correlation is most appropriate as a summary measure of association when the sample (x, y) -values are a random sample of the population. This way, there is a representative sample of the x variation as well as the y variation.

Example 9.12

Does the SAT Predict College GPA? Consider the association between x = score on the SAT college entrance exam and y = college GPA at end of second year of college. The strength of the correlation depends on the variability in SAT scores in the sample. If we study the association only for students at Harvard University, the correlation will probably be weak, because the sample SAT scores will be concentrated very narrowly at the upper end of the scale. By contrast, if we could randomly sample from the population of *all* high school students who take the SAT and then place those students in the Harvard environment, students with poor SAT scores would tend to have low GPAs at Harvard. We would then observe a much stronger correlation. ■

Other aspects of regression, such as fitting a prediction equation to the data and making inferences about the slope, remain valid when we randomly sample y within a restricted range of x -values. We simply limit our predictions to that range. The slope of the prediction equation is not affected by a restriction in the range of x . For Figure 9.17, for instance, the sample slope equals 0.97 for the full data and 0.96 for the restricted middle set. The correlation makes most sense, however, when both x and y are random, rather than only y .

EXTRAPOLATION IS DANGEROUS

It is dangerous to apply a prediction equation to values of x outside the range of observed values. The relationship might be far from linear outside that range. We may get poor or even absurd predictions by extrapolating beyond the observed range.

To illustrate, the prediction equation $\hat{y} = -0.86 + 0.58x$ in Section 9.2 relating x = poverty rate to y = murder rate was based on observed poverty rates between 8.0 and 26.4. It is not valid to extrapolate much below or above this range. The predicted murder rate for a poverty rate of $x = 0\%$ is $\hat{y} = -0.86$. This is an impossible value for murder rate, which cannot be negative.

Here is another type of inappropriate extrapolation: x being positively correlated with y and y being positively correlated with z does not imply that x is positively correlated with z . For example,⁵ in the United States wealthier people tend to reside in wealthier states and wealthier states tend to have a higher percentage favoring the Democratic candidate in presidential elections, yet wealthier people tend to be *less* likely to vote Democratic.

REGRESSION MODEL WITH ERROR TERMS*

Recall that at each fixed value of x , the regression model permits values of y to fluctuate around their mean, $E(y) = \alpha + \beta x$. Any one observation may fall above that

⁵ See *Red State, Blue State, Rich State, Poor State* by A. Gelman (Princeton University Press, 2008).

mean (i.e., above the regression line) or below that mean (i.e., below the regression line). The standard deviation σ summarizes the typical sizes of the deviations from the mean.

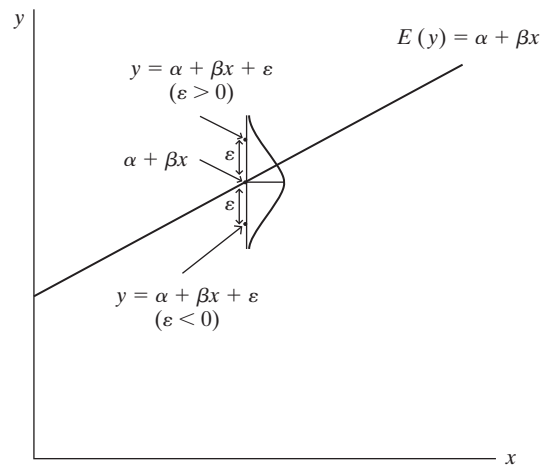
An alternative formulation for the model expresses each observation on y , rather than the mean $E(y)$ of the values, in terms of x . We've seen that the deterministic model $y = \alpha + \beta x$ is unrealistic, because of not allowing variability of y -values. To allow variability, we include a term for the deviation of the observation y from the mean,

$$y = \alpha + \beta x + \varepsilon.$$

The term denoted by ε (the Greek letter epsilon) represents the deviation of y from the mean, $\alpha + \beta x$. Each observation has its own value for ε .

If ε is positive, then $\alpha + \beta x + \varepsilon$ is larger than $\alpha + \beta x$, and the observation falls above the mean. See Figure 9.18. If ε is negative, the observation falls below the mean. When $\varepsilon = 0$, the observation falls exactly at the mean. The mean of the ε -values is 0.

FIGURE 9.18: Positive and Negative ε -Values Correspond to Observations above and below the Mean of the Conditional Distribution



For each x , variability in the y -values corresponds to variability in ε . The ε term is called the **error term**, since it represents the error that results from using the mean value ($\alpha + \beta x$) of y at a certain value of x to predict the individual observation.

In practice, we do not know the n values for ε , just like we do not know the parameter values and the true mean $\alpha + \beta x$. For the sample data and their prediction equation, let e be such that

$$y = a + bx + e.$$

That is, $y = \hat{y} + e$, so $e = y - \hat{y}$. Then e is the **residual**, the difference between the observed and predicted values of y , which we *can* observe. Since $y = \alpha + \beta x + \varepsilon$, the residual e estimates ε . We can interpret ε as a **population residual**. Thus, ε is the difference between the observation y and the mean $\alpha + \beta x$ of all possible observations on y at that value of x . Graphically, ε is the vertical distance between the observed point and the true regression line.

In summary, we can express the regression model either as

$$E(y) = \alpha + \beta x \quad \text{or as} \quad y = \alpha + \beta x + \varepsilon.$$

We use the first equation in later chapters, because it connects better with regression models for response variables assumed to have distributions other than the normal. Models for discrete quantitative variables and models for categorical variables are expressed in terms of their means, not in terms of y itself.

MODELS, REALITY, AND ALTERNATIVE APPROACHES

We emphasize again that the regression model *approximates* the true relationship. No sensible researcher expects a relationship to be exactly linear, with exactly normal conditional distributions at each x and with exactly the same standard deviation of y -values at each x -value. Models merely approximate reality.

If the model seems too simple to be adequate, the scatterplot or other diagnostics may suggest improvement by using other models introduced later in this text. Such models can be fitted, rechecked, and perhaps modified further. Model building is an iterative process. Its goals are to find a realistic model that is adequate for describing the relationship and making predictions but that is still simple enough to interpret easily. Chapters 11–15 extend the basic regression model so that it applies to situations in which the assumptions of this chapter are too simplistic.

9.7 Chapter Summary

Chapters 7–9 have dealt with the detection and description of *association between two variables*. Chapter 7 showed how to compare means or proportions for two groups. When the variables are statistically independent, the population means or proportions are identical for the two groups. Chapter 8 dealt with *association between two categorical variables*. Measures of association such as the difference of proportions, the odds ratio, and gamma describe the strength of association. The chi-squared statistic for nominal data or a z statistic based on sample gamma for ordinal data tests the hypothesis of independence.

This chapter dealt with *association between quantitative variables*. A new element studied here was a regression model to describe the *form* of the relationship between the explanatory variable x and the mean $E(y)$ of the response variable. The major aspects of the analysis are as follows:

- The **linear regression equation** $E(y) = \alpha + \beta x$ describes the *form* of the relationship. This equation is appropriate when a straight line approximates the relationship between x and the mean of y .
- A **scatterplot** views the data and checks whether the relationship is approximately linear. If it is, the **least squares** estimates of the y -intercept α and the slope β provide the prediction equation $\hat{y} = a + bx$ closest to the data, minimizing the sum of squared residuals.
- The **correlation r** and its square describe the *strength* of the linear association. The correlation is a standardized slope, having the same sign as the slope but falling between -1 and $+1$. Its square, r^2 , gives the proportional reduction in variability about the prediction equation compared to the variability about \bar{y} .
- For inference about the relationship, a t test using the slope or correlation tests the **null hypothesis of independence**, namely, that the population slope and correlation equal 0. A confidence interval estimates the size of the effect.

Table 9.9 summarizes the methods studied in the past three chapters.

Chapter 11 introduces the **multiple regression** model, a generalization that permits *several* explanatory variables in the model. Chapter 12 shows how to include categorical predictors in a regression model. Chapter 13 includes both categorical and quantitative predictors. Chapter 14 introduces models for more complex relationships, such as nonlinear ones. Finally, Chapter 15 presents regression models for

INTRODUCTION TO MULTIVARIATE RELATIONSHIPS

Chapter 10

CHAPTER OUTLINE

- 10.1 Association and Causality
- 10.2 Controlling for Other Variables
- 10.3 Types of Multivariate Relationships
- 10.4 Inferential Issues in Statistical Control
- 10.5 Chapter Summary

Chapters 7–9 introduced methods for analyzing the association between two variables. In most social science research, these analyses are but the first step. Subsequent steps use *multivariate* methods to include other variables in the analysis that might influence that association.

For instance, Examples 8.1 and 8.3 showed that political party identification in the United States is associated with gender, with men somewhat more likely than women to be Republicans. To analyze why this is so, we could analyze whether differences between men and women in political ideology (measured on a conservative–liberal scale) could explain the association. For example, perhaps men tend to be more conservative than women, and being conservative tends to be associated with being Republican. If we compare men to women just for those classified as liberal in political ideology, and then again just for those classified as conservative, is it still true that men are more likely than women to be Republicans? Or, could the difference between men and women on political party ID be explained by some other factor, such as income or religion?

Several types of research questions require adding variables to the analysis. These questions often involve notions of *causal* connections among the variables. This chapter discusses causation and outlines methods for testing causal hypotheses. We present the notion of *statistical control*, a fundamental tool for studying how an association changes or possibly even disappears after we remove the influence of other variables. We also show the types of multivariate relationships that statistical control can reveal.

10.1 Association and Causality

Causality is central to the scientific endeavor. Most people are familiar with this concept, at least in an informal sense. We know, for instance, that being exposed to a virus can cause the flu and that smoking can cause lung cancer. But how can we judge whether there is a causal relationship between two social science variables? For instance, what causes juvenile delinquency? Being poor? Coming from a single-parent home? A lack of moral and religious training? Genetic factors? A combination of these and other factors? We now look at some guidelines that help us assess a hypothesis of the form “*x* causes *y*.”

Causal relationships usually have an asymmetry, with one variable having an influence on the other, but not vice versa. An arrow drawn between two variables *x* and *y*, pointing to the response variable, denotes a causal association between the variables. Thus,

$$x \rightarrow y$$

specifies that *x* is an explanatory variable having a causal influence on *y*. For example, suppose we suspect that being a Boy Scout has a causal effect on being a juvenile delinquent, scouts being less likely to be delinquents. We are hypothesizing that

$S \rightarrow D$, where S (for Scouting) and D (for Delinquency) denote the binary variables “whether a Boy Scout (yes, no)” and “whether a juvenile delinquent (yes, no).”

If we suspect that one variable is causally explained by another, how do we analyze whether it actually is? A relationship must satisfy three criteria to be considered a causal one. These criteria, which we’ll discuss below, are

- association between the variables,
- an appropriate time order, and
- the elimination of alternative explanations.

If all three are met, then the evidence supports the hypothesized causal relationship. If one or more criteria are not met, then we conclude that there is not a causal relationship.

ASSOCIATION IS REQUIRED, BUT NOT SUFFICIENT, FOR CAUSALITY

The first criterion for causality is **association**. We must show that x and y are associated. If $x \rightarrow y$, then as x changes, the distribution of y should change in some way. If scouting causes lower delinquency rates, for example, then the population proportion of delinquents should be higher for nonscouts than for scouts. For sample data, a statistical test, such as chi-squared for categorical data or a t test for the regression slope or for a comparison of means for quantitative data, analyzes whether this criterion is satisfied.

Association by itself cannot establish causality.

Association does not imply causation.

The remainder of this section explains why.

CAUSALITY REQUIRES APPROPRIATE TIME ORDER

The second criterion for causality is that the two variables have the appropriate **time order**, with the cause preceding the effect. Sometimes this is just a matter of logic. For instance, race, age, and gender exist prior to current attitudes or achievements, so any causal association must treat them as causes rather than effects.

In other cases, the causal direction is not as obvious. Consider scouting and delinquency. It is logically possible that scouting reduces delinquency tendencies. On the other hand, it is also possible that delinquent boys avoid scouting but nondelinquent boys do not. Thus, the time order is not clear, and both possibilities, $S \rightarrow D$ and $D \rightarrow S$, are plausible. Just showing that an association exists does not solve this dilemma, because a lower proportion of delinquents among scout members is consistent with both explanations.

It is difficult to study cause and effect when two variables do not have a time order but are measured together over time. The variables may be associated merely because they both have a time trend. For example, for recent annual data there is a correlation of 0.993 between y = divorce rate in Maine and x = per capita consumption of margarine.¹ They both have a decreasing trend over time, so they have a strong positive correlation, with higher divorce rates occurring in years that have higher consumption of margarine. Each variable would be strongly negatively correlated with all variables that have a positive time trend, such as percentage of people who

¹ See www.tylervigen.com/spurious-correlations.

use smart phones, percentage of people who belong to an Internet social network such as Facebook, and annual average worldwide temperature.

ALTERNATIVE EXPLANATION MAY INVALIDATE CAUSAL RELATION

When two variables are associated and have the proper time order to satisfy a casual relation, this is still insufficient to imply causality. The association may have an *alternative explanation*.

For example, airline pilots turn on the “fasten seat belt” sign just before their planes encounter turbulence. We observe an association, greater turbulence occurring when the sign is on than when it is off. There is usually also the appropriate time order, the sign coming on, followed by turbulence shortly afterward. But this does not imply that turning on the sign causes turbulence.

An alternative explanation for an association is responsible for rejecting many hypotheses of causal relationships. Many alternative explanations involve an additional variable z or a set of variables. For example, there may be a variable z that causes both x and y .

With observational data, it is easy to find associations, but those associations are often explained by other variables that may not have been measured in a study. For example, some medical studies have found associations between coffee drinking and various responses, such as the likelihood of a heart attack. But after taking into account other variables associated with the extent of coffee drinking, such as country of residence, occupation, and levels of stress, such associations have disappeared or weakened considerably.

This criterion for causality of eliminating an alternative explanation is the most difficult to achieve. We may think we’ve found a causal relationship, but we may merely have not thought of a particular reason that can explain the association. Because of this, *with observational studies we can never prove that one variable is a cause of another*. We can disprove causal hypotheses, however, by showing that empirical evidence contradicts at least one of these three criteria.

ASSOCIATION, CAUSALITY, AND ANECDOTAL EVIDENCE

The association between smoking and lung cancer is regarded as having a causal link. The association is moderately strong, there is the proper time order (lung cancer following a period of smoking), and no alternative explanation has been found to explain the relationship. In addition, the causal link has been bolstered by biological theories that explain how smoking could cause lung cancer.

Sometimes you hear people give anecdotal evidence to attempt to disprove causal relationships. “My Uncle John is 85 years old, he still smokes a pack of cigarettes a day, and he’s as healthy as a horse.” An association does not need to be perfect, however, to be causal. Not all people who smoke two packs of cigarettes a day will get lung cancer, but a much higher percentage of them will do so than people who are nonsmokers. Perhaps Uncle John is still in fine health, but that should not encourage us to tempt the fates by smoking a pack each day. Anecdotal evidence is not enough to disprove causality unless it can deflate one of the three criteria for causality.

ESTABLISHING CAUSALITY WITH RANDOMIZED EXPERIMENTS

As mentioned in Section 2.2 (page 14), a randomized experiment is the ideal way to compare two groups. This approach, by which we randomly select the subjects

for each group and then observe the response, provides the gold standard for establishing causality. For instance, does a new drug have a beneficial effect in treating a disease? We could randomly assign subjects suffering from the disease to receive either the drug or a placebo. Then, to analyze whether the drug assignment may have a causal influence on the response outcome, we would observe whether the proportion successfully treated was significantly higher for the drug group.

In a randomized experiment, suppose that we observe an association between the group variable and the response variable, such as a statistically significant difference between two proportions for a categorical response or between two means for a quantitative response. With such an experiment, we do not expect another variable to provide an alternative explanation for the association. With randomized assignment to groups, the two groups should have about the same distributions for variables not observed but which may be associated with the response variable. So, the association is not because of an association between the group variable and an alternative variable. In addition, when a research study is experimental, the time order is fixed. The outcome for a subject is observed *after* the group is assigned, so the time order is certain. Because of these factors, *it is easier to assess causality with randomized experiments than with observational studies.*

In most social research, unfortunately, randomized experiments are not possible. If we want to investigate the effect of level of education on political ideology, we cannot randomly assign children to different levels of attained education and then later ask them about their ideology. For each person sampled, we can merely observe their actual attained education and political ideology, and data are missing for that subject about what the political ideology would have been had they attained a different level of education. In the next section, we present a way that we can attempt to adjust for different groups differing in their distributions of other variables that could be associated with the response variable.

10.2 Controlling for Other Variables

A fundamental component to evaluating whether x could cause y is searching for an alternative explanation. We do this by studying whether the association between x and y remains when we remove the effects of other variables on this association. In a multivariate analysis, a variable is said to be **controlled** when its influence is removed.

A laboratory experiment controls variables that could affect the results by holding their values constant. For instance, an experiment in chemistry or physics might control temperature and atmospheric pressure by holding them constant in a laboratory environment during the course of the experiment. A lab experiment investigating the effect of different doses of a carcinogen on mice might control the age and diet of the mice.

Randomized experiments cannot strictly control other variables. But in their randomization, we expect groups on which we perform randomization to have similar distributions on the other variables. So, randomized experiments inherently control other variables in a probabilistic sense.

STATISTICAL CONTROL IN SOCIAL RESEARCH

Unlike laboratory sciences, social research is usually observational rather than experimental. We cannot fix values of variables we might like to control, such as intelligence or education, before obtaining data on the variables of interest. But we can approximate an experimental type of control by grouping together observations

with equal, or similar, values on the control variables. Socioeconomic status or a related variable such as education or income is often a prime candidate for control in social research. To control education, for instance, we could group the sample results into those subjects with less than a high school education, those with a high school education but no college education, those with some college education, and those with at least one college degree. This is **statistical control**, rather than experimental control.

The following example, although artificial, illustrates statistical control in a social science setting, by holding a key variable constant.

Example 10.1

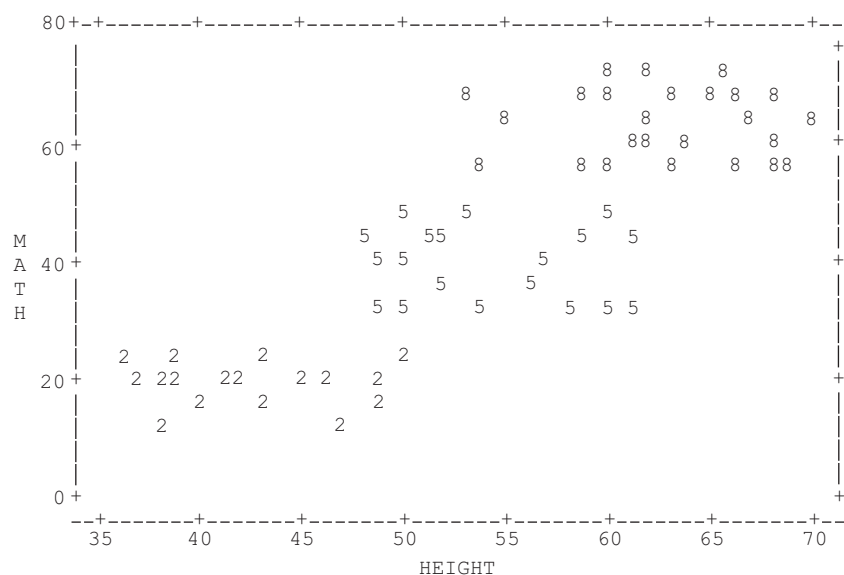
Causal Effect of Height on Math Achievement? Do tall students tend to be better than short students in learning math skills? We might think so looking at a random sample of students from Lake Wobegon school district who take a math achievement test. The correlation is 0.81 between height and math test score. Taller students tend to have higher scores.

Is being tall a causal influence on math achievement? Perhaps an alternative explanation for this association is that the sample has students of various ages. As age increases, both height and math test score would tend to increase. Older students tend to be taller, and older students tend to have stronger math knowledge.

We can remove the effects of age from the association by *statistical control*, studying the association between height and math test score for students of the same age. That is, we control for age by analyzing the association separately at each age level. Then, variation in age cannot jointly cause variation in both height and test score.

In fact, the achievement test was administered to students from grades 2, 5, and 8 at Lake Wobegon, so the sample contained considerable variability in the students' ages. Figure 10.1 shows a scatterplot of the observations, with labels indicating the grade for each student. The overall pattern of points shows a strong positive correlation, with higher math scores at higher heights. View the points within a fixed grade level (for which age is approximately constant), however, and you see random variation, with no particular pattern of increase or decrease. The correlation between height and math test score is close to zero for students of about the same age. Height does not have a causal effect on math test score, because the association disappears when age is held constant. ■

FIGURE 10.1: Scatterplot Showing Relationship between Height and Math Achievement Test Score, with Observations Labeled by Grade Level. Students at a particular grade level have about the same age and show a lack of association between height and test score.



In summary, we control a variable by holding its value constant, or nearly so. We can then study the relationship between x and y for cases with equal, or similar, values of that variable. The variable controlled is called a **control variable**. In holding the control variable constant, we remove the influence of that variable on the association between x and y .

STATISTICAL CONTROL FOR VARIABLE TYPES IN AN ASSOCIATION

The scatterplot in Figure 10.1 describes the association between two quantitative variables, controlling for a third variable. We can describe association between a quantitative variable and a categorical variable by comparing means. For example, at your school suppose the mean salary for male faculty is higher than the mean salary for female faculty. Suppose that the percentage of professors who are female is lowest for full professors and is considerably higher for instructors and assistant professors, perhaps because relatively few female faculty were hired until recent years. Then, this difference in mean salaries would diminish and could even disappear if we control for academic rank.

To study the association between two categorical variables, while controlling for a third variable, we form contingency tables relating those variables separately for subjects at each level of that control variable. The separate tables that display the relationships within the fixed levels of the control variable are called **partial tables**.

Example
10.2

Partial Tables for Control with Categorical Variables Table 10.1 is a hypothetical table relating scouting to delinquency. The percentage of delinquents among scout members is lower than among nonscouts. This table is **bivariate**, meaning that it contains data only on *two* variables. All other variables are ignored. None is controlled.

TABLE 10.1: Contingency Table Relating Scouting and Delinquency. Percentages refer to conditional distribution of delinquency, given whether a boy scout.						
		Delinquency				Total
		Yes		No		
Boy Scout	Yes	36	(9%)	364	(91%)	400
	No	60	(15%)	340	(85%)	400

In seeking a possible explanation for the association, we could control for church attendance. Perhaps boys who attend church are more likely than nonattenders to be scouts, and perhaps boys who attend church are less likely to be delinquent. Then, the difference in delinquency rates between scouts and nonscouts might be due to variation in church attendance.

To control for church attendance, we examine the association between scouting and delinquency within partial tables formed by various levels of church attendance. Table 10.2 shows partial tables for three levels: Low = no more than once a year, Medium = more than once a year but less than once a week, and High = at least once a week. Adding these three partial tables together produces the bivariate table (Table 10.1), which ignores church attendance. For instance, the number of Boy Scouts who are delinquents is $36 = 10 + 18 + 8$.

TABLE 10.2: Contingency Table Relating Scouting and Delinquency, Controlling for Church Attendance

Delinquency	Church Attendance					
	Low		Medium		High	
	Yes	No	Yes	No	Yes	No
Scout	10 (20%)	40 (80%)	18 (12%)	132 (88%)	8 (4%)	192 (96%)
Not scout	40 (20%)	160 (80%)	18 (12%)	132 (88%)	2 (4%)	48 (96%)

In each partial table, the percentage of delinquents is the same for scouts as for nonscouts. Controlling for church attendance, no association appears between scouting and delinquency. These data provide an alternative explanation for the association between scouting and delinquency, making us skeptical of any causal links. The alternative explanation is that both these variables are associated with church attendance. Youngsters who attend church are less likely to be delinquents and more likely to be scouts. For a fixed level of church attendance, scouting has no association with delinquency. Since the association can be explained by church attendance, no causal link exists between scouting and delinquency. ■

Some examples in this chapter, like this one, use artificial data in order to make it simpler to explain the concepts. In practice, some distortion occurs because of sampling variation. Even if an association between two variables truly disappears under a control, *sample* partial tables would not look exactly like those in Table 10.2. Because of sampling variation, they would not show a *complete* lack of association. Moreover, few associations disappear *completely* under a control. There may be *some* causal connection between two variables, but not as strong as the bivariate table suggests. Moreover, in practice we need to control for several variables, and we'll see in the next chapter that statistical control then involves further assumptions.

BE WARY OF LURKING VARIABLES

It is not always obvious which variables require control in a study. Knowing about the theory and previous research in a field of study helps a researcher to know which variables to control. A potential pitfall of almost all social science research is the possibility that the study did not include an important variable. If you fail to control for a variable that strongly influences the association between the variables of primary interest, you will obtain misleading results.

A variable that is *not* measured in a study (or perhaps even known about to the researchers) but that influences the association under study is called a ***lurking variable***. In analyzing the positive correlation between height and math achievement in Example 10.1 (page 291), we observed that the correlation could be due to a lurking variable, the age of the student.

When you read about a study that reports an association, try to think of a lurking variable that could be responsible. For example, suppose a study reports a positive correlation between individuals' college GPA and their income later in life. Is doing well in school responsible for the higher income? An alternative explanation is that both high GPA and high income could be caused by a lurking variable such as IQ or an individual's tendency to work hard.

10.3 Types of Multivariate Relationships

Section 10.2 showed that an association may change dramatically when we control for another variable. This section describes types of multivariate relationships that often occur in social science research. For a response variable y , there may be several explanatory variables and control variables, and we denote them by x_1, x_2, \dots .

SPURIOUS ASSOCIATIONS

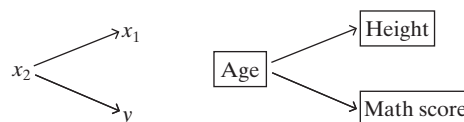
An association between y and x_1 is said to be *spurious* if both variables are dependent on a third variable x_2 , and their association disappears when x_2 is controlled. Such an association results from the relationship of y and x_1 with the control variable x_2 , rather than indicating a causal connection. Showing that the association between two variables may be spurious provides an alternative explanation to a causal connection between them.

Example 10.3

Examples of Spurious Associations Table 10.1 (page 292) displayed an association between scouting and delinquency. Controlling for church attendance, the partial tables in Table 10.2 (page 293) showed no association. This is consistent with spuriousness. Table 10.2 shows that as church attendance increases, the percentage of delinquents decreases (compare percentages across the partial tables) and the percentage of scout members increases. By the nature of these two associations, it is not surprising that Table 10.1 exhibits lower overall delinquency rates for scouts than nonscouts.

The association between height and mathematics achievement test score in Example 10.1 disappears at fixed levels of age. That association is spurious, with age being a common cause of both height and math achievement. Figure 10.2 graphically depicts this spurious association, using $x_1 = \text{height}$ and $y = \text{math test score}$. They are associated only because they both depend on a common cause, $x_2 = \text{age}$. As x_2 changes, it produces changes simultaneously in x_1 and y , so that x_1 and y are associated. In fact, they are associated only because of their common dependence on the third variable (age). ■

FIGURE 10.2: Graphical Depiction of a Spurious Association between y and x_1 . The association disappears when we control for x_2 , which causally affects both x_1 and y .



Example 10.4

Do Fewer Vacations Cause Increased Risk of Death? When an association is observed between two variables, later studies often attempt to determine whether that association might be spurious, by controlling for variables that could be a common cause. For example, some studies have observed an association between frequency of vacationing and quality of health. In particular, a study using a 20-year follow-up of women participants in the Framingham Heart Study found² that less frequent vacationing was associated with greater frequency of deaths from heart attacks.

² E. D. Eaker et al., *American Journal of Epidemiology*, vol. 135 (1992), pp. 835–864.

A later study³ questioned whether this could be a spurious association, explained by the effects of socioeconomic status (SES). For example, perhaps higher SES is responsible both for lower mortality and for more frequent vacations. But after controlling for education, family income, and other potentially important variables with a much larger data set, this study also observed higher risk of heart disease and related death for those who took less vacation time. Perhaps the association is not spurious, unless researchers find another variable to control such that the association disappears. ■

CHAIN RELATIONSHIPS AND INTERVENING (MEDIATOR) VARIABLES

Another way that an association can disappear when we control for a third variable is with a *chain* of causation, in which x_1 affects x_2 , which in turn affects y . Figure 10.3 depicts the chain. Here, x_1 is an *indirect*, rather than direct, cause of y . Variable x_2 is called an *intervening* variable or a *mediator* variable.

FIGURE 10.3: A Chain Relationship, in Which x_1 Indirectly Affects y through an Intervening Variable x_2 , Which Has a Mediating Effect

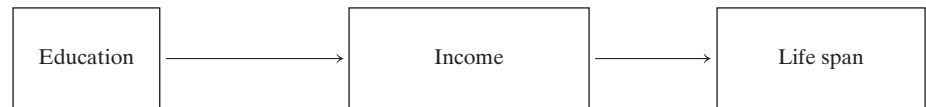
$$x_1 \longrightarrow x_2 \longrightarrow y$$

Example 10.5

Is Education Responsible for a Long Life? A *New York Times* article⁴ summarized research studies dealing with human longevity. It noted that consistently across studies in many nations, life span was positively associated with educational attainment. Many researchers believe education is the most important variable in explaining how long a person lives. Is having more education responsible for having a longer life?

Establishing causal connections is difficult. In some societies, perhaps the causation could go in the other direction, with sick children not going to school or dropping out early because they were ill. Many researchers believe there could be a chain of causation, perhaps with income as an intervening variable. For example, perhaps having more education leads to greater wealth, which then (possibly for a variety of reasons, such as access to better health care) leads to living longer. Figure 10.4 depicts this causal chain model.

FIGURE 10.4: Example of a Chain Relationship. Income is an intervening variable (also called a *mediator* variable), and the association between education and life span disappears when it is controlled.



This model is supported if the association between education and life span disappears after controlling for income; that is, if within fixed levels of income (the intervening variable), no significant association occurs. If this happens, education does not directly affect life span, but it is an indirect cause through income. ■

For both spurious relationships and chain relationships, an association between y and x_1 disappears when we control for a third variable, x_2 . The difference between

³ B. B. Gump and K. A. Matthews, *Psychosomatic Medicine*, vol. 62 (2000), pp. 608–612.

⁴ Written by G. Kolata, January 3, 2007.

the two is in the causal order among the variables. For a spurious association, x_2 is causally prior to both x_1 and y , as in Figure 10.2. In a chain association, x_2 intervenes between the two, as in Figures 10.3 and 10.4.

To illustrate, a study⁵ of mortality rates in the United States found that states that had more income inequality tended to have higher age-adjusted mortality rates. However, this association disappeared after controlling for the percentage of a state's residents who had at least a high school education. Might this reflect a chain relationship, or a spurious relationship? Greater education could tend to result in less income inequality, which could in turn tend to result in lower mortality rates. Thus, the chain relationship

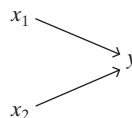
Education \longrightarrow Income inequality \longrightarrow Mortality rate

is plausible. For the relationship to be spurious, education would need to have a causal effect on both income inequality and mortality. This is also plausible. Just from viewing the association patterns, we do not know which provides a better explanation.

MULTIPLE CAUSES

Response variables in social science research almost always have more than one cause. For instance, a variety of factors likely have causal influences on responses such as y = juvenile delinquency or y = length of life. Figure 10.5 depicts x_1 and x_2 as separate causes of y . We say that y has **multiple causes**.

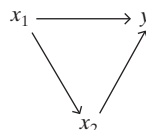
FIGURE 10.5: Graphical Depiction of Multiple Causes of y



Sometimes variables that are separate causes of y are themselves statistically independent. That is, they are *independent causes*. For instance, x_1 = gender and x_2 = race are essentially statistically independent. If they both have effects on juvenile delinquency, with delinquency rates varying both according to gender and race, they are likely to be independent causes.

In the social sciences, most explanatory variables are associated. Both being poor and being from a single-parent family may cause delinquency, but those factors are themselves probably associated. Because of complex association linkages, when we control for a variable x_2 or a set of variables x_2, x_3, \dots , the x_1y association usually changes somewhat. Often the association decreases somewhat, although usually it does not completely disappear as in a spurious or chain relationship. Sometimes this is because x_1 has direct effects on y and also indirect effects through other variables. Figure 10.6 illustrates this. For instance, perhaps being from a single-parent family has direct effects on delinquency but also indirect effects through being more likely to be poor. Most response variables have many causes, both direct and indirect.

FIGURE 10.6: Graphical Depiction of Direct and Indirect Effects of x_1 on y



⁵ A. Muller, *BMJ*, vol. 324 (2002), pp. 23–25.

SUPPRESSOR VARIABLES

In examples so far, an association disappears or weakens when we control for another variable. By contrast, occasionally two variables show no association until a third variable is controlled. That control variable is called a *suppressor variable*.

Example 10.6

Age Suppresses the Association between Education and Income Is educational level positively related with income? Table 10.3 shows such a relationship, measured as binary variables, controlling for age. In each partial table, the percentage of subjects at the high level of income is greater when education is high than when education is low.

TABLE 10.3: Partial Tables Relating Education and Income, Controlling for Age

		Age = Low			Age = High		
Income:		High	Low	% High	High	Low	% High
Education	High	125	225	35.7	125	25	83.3
	Low	25	125	16.7	225	125	64.3

Suppose now that we ignore age, adding these two partial tables together. The bivariate table for education and income is the first panel of Table 10.4. Every count equals 250. Both when education is high and when education is low, the percentage having a high income is 50%. For the bivariate table, no association exists between education and income.

TABLE 10.4: Bivariate Tables Relating Education, Income, and Age

Education	Income		Age	Income		Age	Education	
	High	Low		High	Low		High	Low
High	250	250	High	350	150	High	150	350
Low	250	250	Low	150	350	Low	350	150

A look at the other two bivariate tables in Table 10.4 reveals how this could happen. Age is positively associated with income but negatively associated with education. Older subjects tend to have higher income, but they tend to have lower education. Thus, when we ignore rather than control age, we give an inadvertent boost to the relative numbers of people at high incomes with low educational levels and at low incomes with high educational levels. Because of the potential for a suppressor effect, it can be informative to control for a variable even when the bivariate analysis does not show an association with y . This is especially true when there is a theoretical reason for a potential suppression effect. ■

STATISTICAL INTERACTION

Often the effect of an explanatory variable on a response variable changes according to the level of another explanatory variable or control variable. When the true effect of x_1 on y changes at different levels of x_2 , the relationship is said to exhibit *statistical interaction*.

Statistical Interaction

Statistical interaction exists between x_1 and x_2 in their effects on y when the true effect of one predictor on y changes as the value of the other predictor changes.

**Example
10.7**

Interaction between Education and Gender in Predicting Income Consider the relationship between y = annual income (in thousands of dollars) and x_1 = number of years of education, by x_2 = gender. Many studies in the United States have found that the slope for a regression equation relating y to x_1 is larger for men than for women. Suppose that in the population, the regression equations are $E(y) = -10 + 5x_1$ for men and $E(y) = -5 + 3x_1$ for women. On the average, income for men increases by \$5000 for every year of education, whereas for women it increases by \$3000 for every year of education. That is, the effect of education on income varies according to gender, with the effect being greater for men than for women. So, there is interaction between education and gender in their effects on income. ■

**Example
10.8**

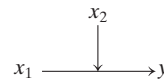
Interaction between SES and Age in Predicting Health Some studies⁶ have noted that quality of health (measured by self-rated and health indexes) tends to be positively associated with SES (measured by years of education and annual household income), and that the association strengthens with age. For example, the gap in health between low SES and high SES levels tends to be larger at older ages. Thus, there is interaction between SES and age in their effects on health. ■

ANALYZING AND DEPICTING INTERACTION

To assess whether a sample shows evidence of interaction, we can compare the effect of x_1 on y at different levels of x_2 . When the sample effect is similar at each level of x_2 , it's simpler to use statistical analyses that assume an absence of interaction. The interaction is worth noting when the variability in effects is large. For instance, perhaps the association is positive at one level of x_2 and negative at another, or strong at one level and weak or nonexistent at another.

Figure 10.7 depicts a three-variable relationship having statistical interaction. Here, x_2 affects the relationship between x_1 and y . When this happens, then likewise x_1 affects the relationship between x_2 and y .

FIGURE 10.7: Graphical Depiction of Statistical Interaction. The effect of one explanatory variable on y depends on the level of the other explanatory variable.



Suppose *no* interaction occurs between x_1 and x_2 in their effects on y . This does not mean that x_1 and x_2 have no association. There can be a lack of statistical interaction even when all the variables are associated. For instance, Tables 10.2 (page 293) and 10.3 (page 297) showed no interaction—in each case the association was similar in each partial table. However, in each case the explanatory variables

⁶ For example, see S. G. Prus, *Canadian Journal on Aging*, vol. 23 (2004), Supplement, pp. S145–S153.

were associated, with each other and with the response. In Table 10.3, for instance, age was negatively associated with education and positively associated with income.

SUMMARY OF MULTIVARIATE RELATIONSHIPS

In summary,

- For spurious relationships (i.e., x_2 affects both x_1 and y) and chain relationships (i.e., x_2 intervenes between x_1 and y), the x_1y association disappears when we control for x_2 .
- For multiple causes, an association may change under a control but does not disappear.
- When there is a suppressor variable, an association appears only under the control.
- When there is statistical interaction, an association has different strengths and/or directions at different values of a control variable.

This does not exhaust the possible association structures. It is even possible that, after controlling for a variable, each association in a partial table has the *opposite* direction as the bivariate association. This is called **Simpson's paradox** and is illustrated in Exercises 10.14, 10.29, and 10.30.

CONFOUNDING AND OMITTED VARIABLE BIAS

When two explanatory variables both have effects on a response variable but are also associated with each other, there is said to be **confounding**. It is difficult to determine whether either of them truly causes the response, because a variable's effect could be at least partly due to its association with the other variable. We usually observe a different effect on y for a variable when we control for the other variable than when we ignore it.

In analyzing the effect of an explanatory variable of key interest, if our study neglects to observe a confounding variable that explains a major part of that effect, our results and conclusions will be biased. Such bias is called **omitted variable bias**.

Confounding and omitted variable bias are constant worries in social science research. They are the main reason it is difficult to study many issues of importance, such as what causes crime or what causes the economy to improve or what causes students to succeed in school.

10.4 Inferential Issues in Statistical Control

To conduct research well, you must select the key variables, determine which variables to control, choose an appropriate model, and analyze the data and interpret the results properly. So far this chapter has ignored inferential matters, to avoid confusing them with the new concepts presented. We now discuss some inferential issues in studying associations while controlling other variables.

EFFECTS OF SMALLER SAMPLE SIZE IN PARTIAL ANALYSES

Suppose we control for x_2 in studying the x_1y association. The sample size at a fixed level of x_2 may be much smaller than in the full data set. Even if no reduction in

association occurs relative to the full data, standard errors of parameter estimators tend to be larger. Thus, confidence intervals for those parameters at fixed levels of x_2 tend to be wider, and test statistic values tend to be smaller.

For categorical data, for example, we could compute the Pearson X^2 statistic within a particular partial table to test whether the variables are independent at that level of x_2 . This X^2 -value may be small relative to the X^2 -value for the bivariate x_1y table. This could be due partly to a weaker association, but it could also reflect the reduction in sample size. Section 8.4 showed that larger sample sizes tend to produce larger X^2 -values, for a particular degree of association.

EFFECTS OF CATEGORIZATION IN CONTROLLING A VARIABLE

Categorical control variables (e.g., gender) have the categories as the natural values held constant in partial tables. For ordinal control variables, you should avoid overly crude categorizations. The greater the number of control levels, the more nearly constant the control variable is within each partial table. It is usually advisable to use at least three or four partial tables.

On the other hand, it is preferable not to use more partial tables than needed, because then each one may have a small sample size. Separate estimates may have large standard errors, resulting in imprecise inferences within the partial tables and comparisons of associations between tables. Fortunately, the model-building methods presented in the rest of the text allow us to attempt statistical control and assess patterns of association and interaction without necessarily performing separate analyses at the various combinations of levels of the control variables.

COMPARING AND POOLING MEASURES

It is often useful to compare parameter values describing the effect of an explanatory variable on a response variable at different levels of a control variable. You can construct a confidence interval for a difference between two parameter values in the same way as Chapter 7 showed for a difference of proportions or a difference of means. Suppose that the two sample estimates are based on independent random samples, with standard errors se_1 and se_2 . Then, Section 7.1 noted that the standard error for the difference between the estimates is $\sqrt{(se_1)^2 + (se_2)^2}$. For large random samples, most estimates have approximately normal sampling distributions. Then, a confidence interval for the difference between the parameters is

$$(\text{Estimate}_2 - \text{Estimate}_1) \pm z\sqrt{(se_1)^2 + (se_2)^2}.$$

If the interval does not include 0, the evidence suggests that the parameter values differ.

Example 10.9

Comparing Happiness Associations for Men and Women Is there a difference between men and women in the association between happiness and marital happiness? For recent data from the GSS, the sample value of gamma for a 3×3 table relating these two ordinal variables is 0.674 ($se = 0.0614$, $n = 326$) for males and 0.689 ($se = 0.0599$, $n = 350$) for females.

A 95% confidence interval for the difference between the population values of gamma is

$$(0.689 - 0.674) \pm 1.96\sqrt{(0.0614)^2 + (0.0599)^2}, \quad \text{or} \quad 0.015 \pm 0.168,$$

which is $(-0.153, 0.183)$. It is plausible that the population gamma values are identical. If they are not identical, they seem not to be very different. ■

When the association between two variables is similar in the partial analyses, we can form a measure that summarizes the strength of the association, conditional on the control variable. This is referred to as a measure of **partial association**. The rest of the text shows how to do this in various contexts, using models that handle all the variables at once.

10.5 Chapter Summary


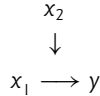

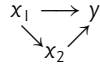
We use a multivariate analysis to study effects of multiple explanatory variables on a response variable. To demonstrate a causal relationship, we must show **association** between variables, ensure proper **time order**, and **eliminate alternative explanations** for the association. This is possible for randomized experiments, but eliminating alternative explanations is a challenge for observational studies.

To consider alternative explanations in observational studies, we introduce **control variables**. We perform statistical control by analyzing associations while keeping the values of control variables essentially constant. This helps us to detect

- **Spuriousness**, in which x_2 jointly affects both y and x_1 .
- **Chain relationships**, in which x_2 is an **intervening variable** (also called a **mediator variable**), so that x_1 affects y indirectly through its effects on x_2 .
- **Suppressor variables**, in which the x_1y association appears only after controlling for x_2 .
- **Statistical interaction**, in which the effect of x_1 on y varies according to the value of x_2 .

Table 10.5 summarizes some possible relationships. The remainder of this text presents statistical methods for multivariate relationships. As you learn about these methods, be careful not to overextend your conclusions: Realize the limitations in making causal inferences with inferential statistical analyses, and keep in mind that any inferences you make must usually be tentative because of assumptions that may be violated or lurking variables that you did not include in your analyses. For further discussion of these points in the context of regression modeling, see Berk (2004), Freedman (2005), Morgan and Winship (2007), and Pedhazur (1997).

TABLE 10.5: Some Three-Variable Relationships

Graph	Name of Relationship	Controlling for x_2
	Spurious x_1y association	Association between x_1 and y disappears.
$x_1 \longrightarrow x_2 \longrightarrow y$	Chain relationship; x_2 intervenes; x_1 indirectly causes y	Association between x_1 and y disappears.
	Interaction	Association between x_1 and y varies according to level of x_2 .
	Multiple causes	Association between x_1 and y does not change.
	Both direct and indirect effects of x_1 on y	Association between x_1 and y changes, but does not disappear.