CHAPTER 9

Linear Regression and Correlation

Chapter 8 presented methods for analyzing association between categorical response and explanatory variables. This chapter presents methods for analyzing quantitative response and explanatory variables.

Table 9.1 shows data from *Statistical Abstract of the United States* for the 50 states and the District of Columbia (D.C.) on the following:

- 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 study methods for analyzing relationships between quantitative variables in this chapter and in some exercises. The text Web site contains two datasets on these and other variables that we will also analyze in exercises in this and later chapters.

We analyze three different, but related, aspects of such relationships:

1. We investigate *whether there is an association* 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. For instance, such an equation predicts a state's murder rate using the percentage of its population living below the poverty level.

The analyses are collectively called a ***regression analysis***. Section 9.1 shows how to use a straight line for the regression equation, and Section 9.2 shows how to use data to estimate the line. Section 9.3 introduces the *linear regression model*, which takes into account variability of the data about the regression line. Section 9.4 uses the *correlation* and its square to describe the strength of association. Section 9.5 presents

TABLE 9.1: Statewide Data 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

statistical inference for a regression analysis. The final section takes a closer look at assumptions and potential pitfalls in using regression.

9.1 LINEAR RELATIONSHIPS

Notation for Response and Explanatory Variables

Let y denote the *response* variable and let x denote the *explanatory* variable.

We shall 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 varies according to the value of x. This formula describes how y = statewide murder varies according to the level of x = percent below the poverty level. Does the murder rate tend to be higher for states that have higher poverty levels?

Linear Functions

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).

EXAMPLE 9.1 Example of a Linear Function

The formula $y = 3 + 2x$ is a linear function. It has the form $y = \alpha + \beta x$ with $\alpha = 3$ and $\beta = 2$. The y-intercept equals 3 and the slope equals 2.

Each real number x, when substituted into the formula $y = 3 + 2x$, yields a distinct value for y. For instance, $x = 0$ has $y = 3 + 2(0) = 3$, and $x = 1$ has $y = 3 + 2(1) = 5$. Figure 9.1 plots this function. 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*. ■

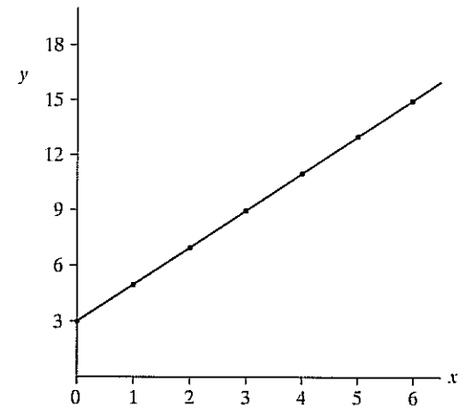

FIGURE 9.1: Graph of the Straight Line $y = 3 + 2x$. The y-intercept is 3 and the slope is 2.

Interpreting the y-Intercept and Slope

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 straight line $y = 3 + 2x$ intersects the y-axis at $\alpha = 3$, as Figure 9.1 shows.

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. For the line $y = 3 + 2x$, $y = 3$ at $x = 0$ and $y = 5$ at $x = 1$. These y values differ by $\beta = 5 - 3 = 2$. Two x-values that are 10 units apart differ by 10β in their y-values. For example, when $x = 0$, $y = 3$, and when $x = 10$, $y = 3 + 2(10) = 23$, and $23 - 3 = 20 = 10\beta$. Figure 9.2 portrays the interpretation of the y-intercept and slope.

To draw the straight line, we find any two separate pairs of (x, y) values on the graph and then draw the line through the points. To illustrate, let's use the points just discussed: $(x = 0, y = 3)$ and $(x = 1, y = 5)$. The point on the graph with $(x = 0, y = 3)$ is three units up the y-axis. To find the point with $(x = 1, y = 5)$, we start at the origin $(x = 0, y = 0)$ and move one unit to the right on the x-axis and five units upward parallel to the y-axis (see Figure 9.1). After plotting the two points, drawing the straight line through the two points graphs the function $y = 3 + 2x$.

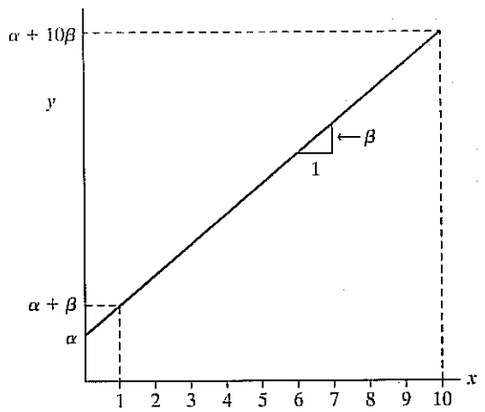

FIGURE 9.2: Graph of the Straight Line $y = \alpha + \beta x$. The y-intercept is α and the slope is β.

EXAMPLE 9.2 Straight Lines for Predicting Violent Crime

For the 50 states, consider the variables y = violent crime rate and x = poverty rate. We'll see that the straight line $y = 210 + 25x$ approximates their relation. The y-intercept equals 210. This 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 predictor. An increase of 1 in the percent below the poverty level corresponds to a greater change in the violent crime rate than an increase of 1 in the percent urban. Figure 9.3 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. ■

If β is positive, then y *increases* as x *increases*—the straight line goes upward, like the two lines just mentioned. Then large values of y occur with large values of x, and small values 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.***

If β is negative, then 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 percent who are high school graduates, the violent crime rate decreases by about 16. Figure 9.3 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 9.3 is an example of a line with $\beta = 0$.

Models Are Simple Approximations for Reality

As Section 7.3 explained, a ***model*** is a simple approximation for the relationship between variables in the population. The linear function is the simplest mathematical

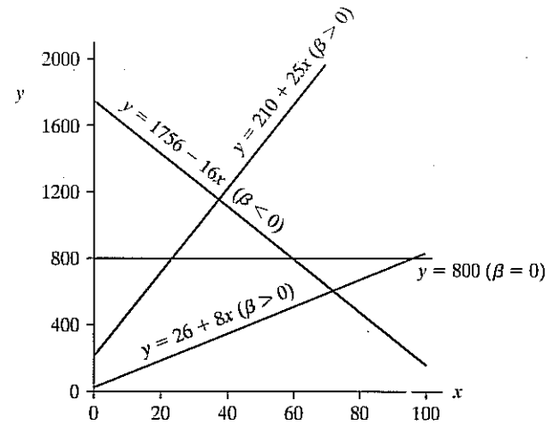

FIGURE 9.3: Graphs of Lines Showing Positive Relationships ($\beta > 0$), a Negative Relationship ($\beta < 0$), and Independence ($\beta = 0$)

function. It provides the simplest 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 mentioned in Section 3.4 and will explain further at the beginning of Chapter 10, *association does not imply causation*. For example, consider the interpretation of the slope in Example 9.2 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 a 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.

9.2 LEAST SQUARES PREDICTION EQUATION

Using sample data, we can estimate the linear model. The process treats α and β in the equation $y = \alpha + \beta x$ as unknown parameters and estimates them. The estimated linear function then provides predicted y-values at fixed values for x.

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 Statewide Murder Rate and Poverty

For Table 9.1, let x = poverty rate and y = murder rate. To check whether a straight line approximates the relationship well, we first construct a scatterplot for the 51 observations. Figure 9.4 shows this plot.

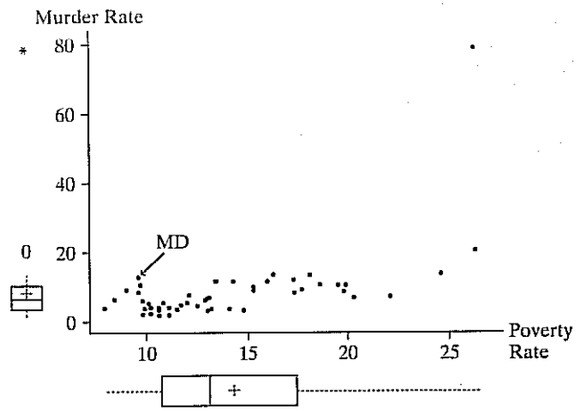

FIGURE 9.4: Scatterplot for y = Murder Rate and x = Percentage of Residents below the Poverty Level, for 50 States and D.C. Box plots are shown for murder rate to the left of the scatterplot and for poverty rate below the scatterplot.

Each point in Figure 9.4 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 9.4.

Figure 9.4 indicates that the trend of points seems to be approximated well by a straight line. Notice, though, that one point is far removed from the rest. This is the point for the District of Columbia (D.C.). For it, the murder rate was much higher than for any state. This point lies far from the overall trend. Figure 9.4 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 9.5 illustrates such a case. This figure shows a negative relationship

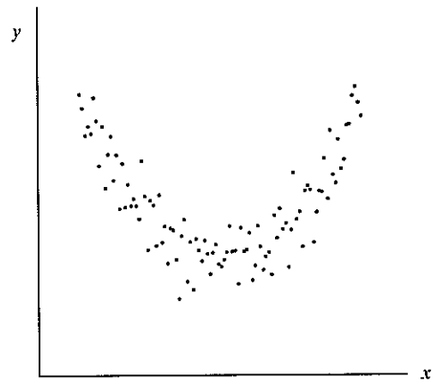

FIGURE 9.5: A Nonlinear Relationship, for Which It Is Inappropriate to Use a Straight Line Model

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 different model, presented in Section 14.5, is appropriate.

Prediction Equation

When the scatterplot suggests that the model $y = \alpha + \beta x$ is realistic, 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. Substituting a particular x-value into $a + bx$ provides a value, denoted by $\hat{y}$, that predicts y at that value of x. 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 discussed later in this section. The formulas for a and b in the prediction equation $\hat{y} = a + bx$ are

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}, \qquad a = \bar{y} - b\bar{x}.$$

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. Figure 9.4 is an example of such a case.

We shall not dwell on these formulas or even illustrate how to use them, as they are messy for hand calculation. Anyone who does any serious regression modeling uses a computer or a calculator that has these formulas programmed. To use statistical software, you supply the data file and usually select the regression method from a menu. The appendix at the end of the text provides details.

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 9.4 shows that the extreme outlying observation for D.C. contributes to this.

TABLE 9.2: Part of SPSS Printout for Fitting Linear Regression Model to Observations for 50 States and D.C. on x = Percent 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 estimates of α and β are listed under the heading "B," the symbol that SPSS uses to denote an estimated regression coefficient. 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 $\hat{y} = a + bx = -10.14 + 1.32x$.

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.

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. Since the mean murder rate is 8.7, it seems that poverty rate is an important predictor of murder rate. 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 9.6 plots the prediction equation from Example 9.4 over the scatterplot. The diagram shows that the observation for D.C. 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.

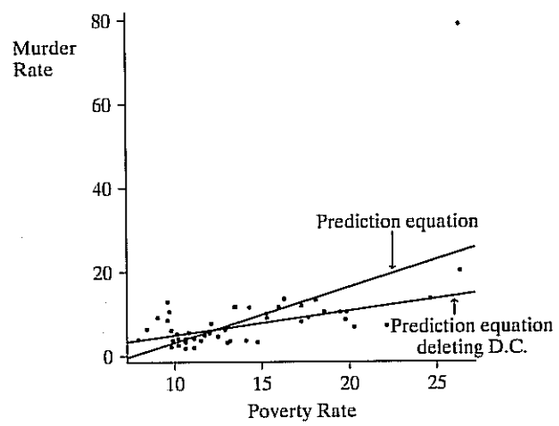

FIGURE 9.6: Prediction Equations Relating Murder Rate and Percentage in Poverty, with and without D.C. Observation

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 equals $\hat{y} = -0.86 + 0.58x$. Figure 9.6 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!

TABLE 9.3: Part of Printout for Fitting Linear Model to 50 States (but not D.C.) on x = Percent in Poverty and y = Murder Rate

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

	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

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 data for the 50 states alone rather than to use a single equation both for 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.

Prediction Errors Are Called Residuals

The prediction equation $\hat{y} = -0.86 + 0.58x$ predicts murder rates using x = poverty rate. For the sample data, a comparison of the *actual* murder rates to the *predicted* values checks the goodness of the prediction equation.

For example, Massachusetts had $x = 10.7$ 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.

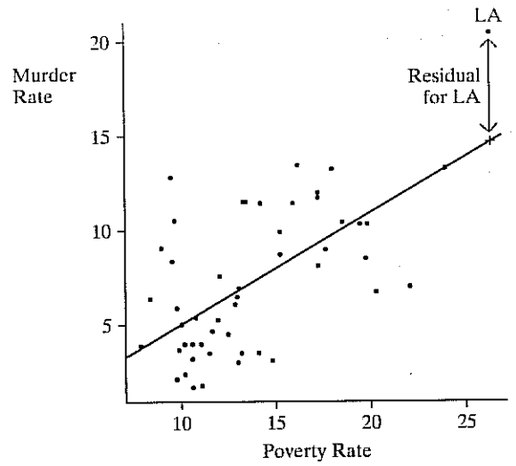

FIGURE 9.7: Prediction Equation and Residuals. A residual is a vertical distance between a point and the prediction line.

In a scatterplot, the residual for an observation is the vertical distance between its point and the prediction line. Figure 9.7 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

Each observation has a residual. If the prediction line falls close to the points in the scatterplot, the residuals are small. We summarize the size of the residuals by the sum of their squared values. This quantity, denoted by SSE, equals

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

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. Some software (such as SPSS) calls SSE the ***residual sum of squares***. It describes the variation of the data around the prediction line.

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

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

This value is smaller than the value of 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.

Software for regression lists the value of SSE. Table 9.3 reports it in the "Sum of Squares" column, in the row labeled "Residual." In some software, such as SAS, this is labeled as "Error" in the sum of squares column.

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

9.3 THE LINEAR REGRESSION MODEL

For the 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 $x = 12$. This distribution refers to the variability in the y values at a *fixed* value of x, so it is a ***conditional distribution***. A separate conditional distribution applies for those with $x = 13$ years of education, and others apply for those with other values of x. 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*, which is another term for the *mean*.

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, rather than stating that *every* subject with $x = 12$ has income 31,000 dollars. The model 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. The y-intercept α and the slope β are called the ***regression coefficients*** for the linear regression function.

In practice, the parameters of the linear regression function are unknown. Least squares provides the sample prediction equation $\hat{y} = a + bx$. At a fixed 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.

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. 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.

The slope $\beta = 3000$ implies that mean income increases \$3000 for each year increase in education. Figure 9.8 shows this regression model. That figure shows the conditional income distributions at $x = 8, 12$, and 16 years. ■

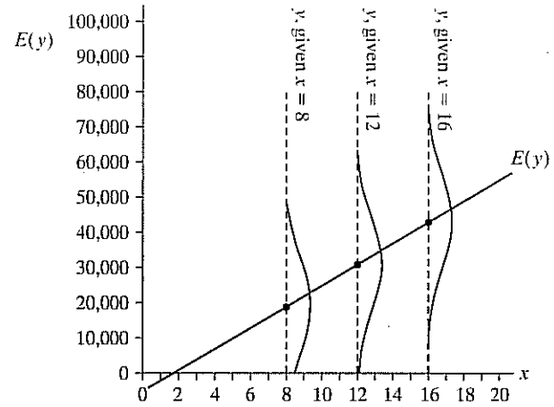

FIGURE 9.8: The Regression Model $E(y) = -5000 + 3000x$, with $\sigma = 13{,}000$, Relating y = Income (in Dollars) to x = Education (in Years)

In Figure 9.8, each conditional distribution is normal, and each has the same standard deviation, $\sigma = 13{,}000$. 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.

Mean Square Error: 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 the numerical value for $\text{SSE} = \sum(y - \hat{y})^2$, which measures sample variability about the least squares line. The estimate is

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

If the constant variation assumption is not valid, then s provides a measure of *average* variability about the line.

EXAMPLE 9.6 TV Watching and Grade Point Averages

A survey[1] 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 $\hat{y} = 3.44 - 0.03x$. For the data, software reports the following:

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

[1] www.iusb.edu/~journal/2002/hershberger/hershberger.html

The residual sum of squares in using x to predict y was SSE = 11.66. The estimated conditional standard deviation is

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

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 $x = 20$, for instance, the conditional distribution of GPA is estimated to have a mean of $3.44 - 0.03(20) = 2.83$ and standard deviation of 0.49. ■

The term $(n - 2)$ in the denominator of s is the ***degrees of freedom*** (df) for the estimate. In general, 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 preceding 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. Some software calls this the MSE, short for *mean square error*. Its square root is the estimate of the conditional standard deviation of y, namely $s = \sqrt{0.24} = 0.49$. (SPSS lists this under the rather misleading heading "Std. Error of the Estimate")

Conditional Variation Tends to be Less than Marginal Variation

From Sections 3.3 and 5.1, a point estimate of the population standard deviation of a variable y is

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

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.

The sum of squares $\Sigma(y - \bar{y})^2$ in the numerator of s_y is called the ***total sum of squares***. In the preceding table 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 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, the *marginal* distribution of college GPAs (y) at your school may primarily fall between 1.0 and 4.0. Perhaps a sample has a standard deviation of $s_y = 0.60$. Suppose we could predict college GPA *perfectly* using x = high school GPA, with the prediction equation $\hat{y} = 0.40 + 0.90x$. Then SSE would be 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 9.9, which portrays a marginal distribution that is much more spread out than each conditional distribution.

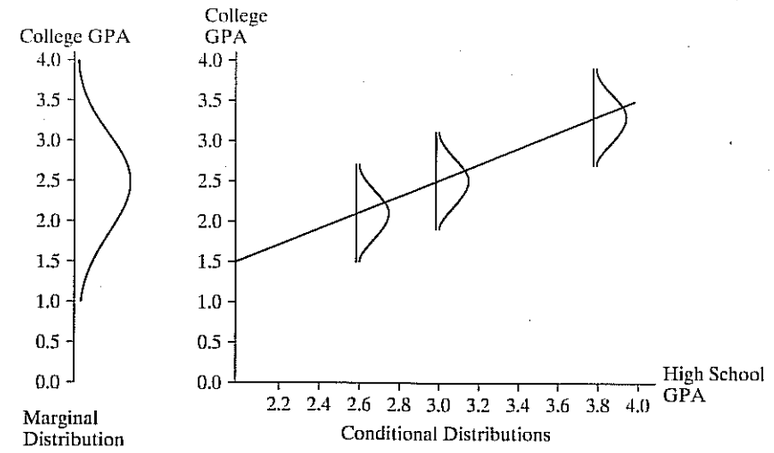

FIGURE 9.9: Marginal and Conditional Distributions. The marginal distribution shows the overall variability in y values, whereas the conditional distribution shows how y varies at a fixed value of x.

9.4 MEASURING LINEAR ASSOCIATION: THE CORRELATION

The linear regression model uses a straight line to describe the relationship. This section introduces two measures of the strength of association between the 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 = percent 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. In summary, 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 percent below the poverty level has a much greater effect on murder rate than an increase of 1 in percent urban.

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

The Correlation

Section 3.5 introduced the ***correlation*** between quantitative variables. This is a *standardized* version of the slope. Its value, unlike that of the ordinary slope b, does not depend on the units of measurement. The standardization adjusts the slope b for the fact that the standard deviations of x and y depend on their units of measurement. The correlation is the value the slope would take for units such that the variables have equal standard deviations.

Let s_x and s_y denote the marginal sample standard deviations of x and y,

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

Correlation

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

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

When the sample spreads are equal ($s_x = s_y$), $r = b$. 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's not necessary to standardize the variables, but it is often useful to 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 for the 50 states 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 and $s_y = 3.98$ for murder rate. The correlation equals

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

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 is a sensible model for the relationship. Since r is proportional to the slope of a linear prediction equation, it measures the *strength of the linear association* between x and y.

- $-1 \leq r \leq 1$. The correlation, unlike the slope b, must fall between -1 and $+1$. The reason will be seen later in the section.
- r has the same sign as the slope b. Since r equals b multiplied by the ratio of two (positive) standard deviations, the sign is preserved. 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 the prediction equation $\hat{y} = a + bx$ predicts 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 9.10 shows scatterplots having various values for r.

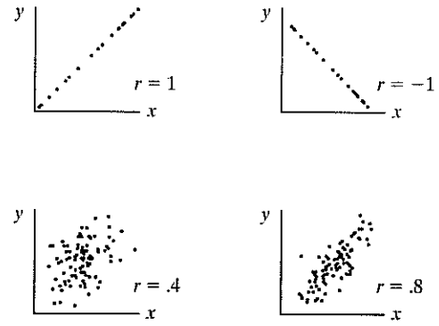

FIGURE 9.10: Scatterplots for Different Correlations

- 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.

Many properties of the correlation are similar to those of the ordinal measure of association *gamma* (Section 8.5). It falls between -1 and $+1$, it is symmetric, and larger absolute values indicate stronger associations.

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 9.11. A low absolute value for r does not then imply that the variables are unassociated, but that the association is not linear.

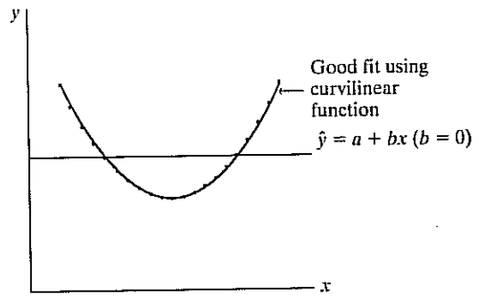

FIGURE 9.11: Scatterplot for Which $r = 0$, Even Though There Is a Strong Curvilinear Relationship

Correlation Implies Regression toward the Mean

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

$$r = (s_x/s_y)b \quad \text{as} \quad s_x b = rs_y.$$

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 9.12. Since $s_x b = rs_y$, an increase of s_x in x corresponds to a predicted change of r standard deviations in the y values. 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.

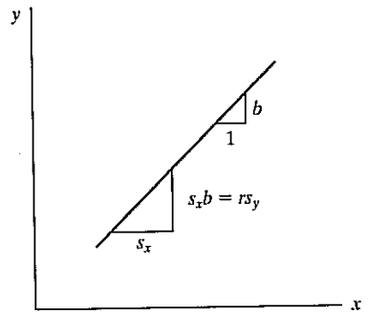

FIGURE 9.12: An Increase of s_x Units in x Corresponds to a Predicted Change of rs_y Units in y

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 (because the correlation is 0.5).

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. ■

For x = poverty rate and y = murder rate for the 50 states, the correlation is $r = 0.63$. So a standard deviation increase in the poverty rate corresponds to a predicted 0.63 standard deviation increase in murder rate. By contrast, $r = 0.37$ between the poverty rate and the violent crime rate. This association is weaker. A standard deviation increase in poverty rate corresponds to a smaller change in the predicted violent crime rate than in the predicted murder rate (in standard deviation units).

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:

- Rule 1 for predicting y without using x.
- Rule 2 for predicting y using information on x.
- 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. For each subject, substituting the x-value into this equation provides the predicted value 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 equals

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

This is the *total sum of squares* of the y-values about their mean. We denote this by TSS. For rule 2, the predicted values are the $\hat{y}$-values from the prediction equation. The total prediction error equals

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

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 9.13 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.

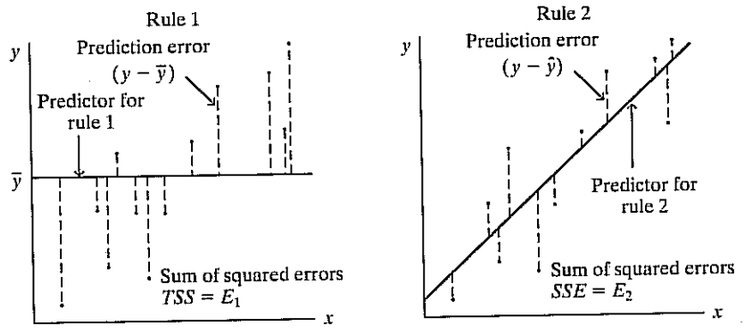

FIGURE 9.13: Graphical Representation of Rule 1 and Total Sum of Squares $E_1 = TSS = \Sigma(y - \bar{y})^2$, Rule 2 and Residual Sum of Squares $E_2 = SSE = \Sigma(y - \hat{y})^2$

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

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

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. We don't need to use the sums of squares in its definition to find r^2, as we can square the correlation. Its defining formula is useful for interpreting r^2, but it is not needed for its calculation.

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 showed

	Sum of Squares
Regression	307.342
Residual	470.406
Total	777.749

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

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

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 $r^2 = TSS/TSS = 1$. For $SSE = 0$, all sample points must fall exactly on the prediction line. In that case, there is no prediction error using x to predict y. 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 that 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 $= \Sigma(y - \bar{y})^2$, 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 $= \Sigma(y - \hat{y})^2$ summarizes the variability around the prediction equation, which refers to variability for the conditional distributions. When $r^2 = 0.39$, the variability in y using x to make the predictions (via the prediction equation) 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.

When two variables are strongly associated, the variation in the conditional distributions is considerably less than the variation in the marginal distribution. Figure 9.9 illustrated this.

9.5 INFERENCES FOR THE SLOPE AND CORRELATION

Sections 9.1–9.3 showed how a linear regression model can represent the *form* of relationships between quantitative variables. Section 9.4 used 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:

- The study used 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 second assumption states that the linear regression function is valid. The assumption about a common σ is one under which the least squares estimates are the best possible estimates of the regression coefficients.[2] 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.

[2]Under the assumptions of normality with common σ, least squares estimates are special cases of *maximum likelihood* estimates, introduced in Section 5.1.

Test of Independence

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 9.14). The null hypothesis that the variables are statistically independent is H_0: $\beta = 0$.

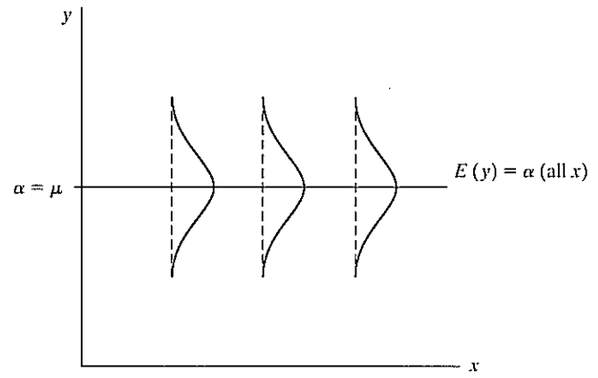

FIGURE 9.14: x and y Are Statistically Independent when the Slope $\beta = 0$ in the Regression Model $E(y) = \alpha + \beta x$

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 equals

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

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 degrees of freedom are the same as the df of the conditional standard deviation estimate s.

The formula for the standard error of b is

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

This depends on the point estimate s of the standard deviation of the conditional distributions of y. 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 $\Sigma(x - \bar{x})^2$, 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 P-value for H_a: $\beta \neq 0$ is the two-tail probability from the t distribution. Software provides the P-value. For large df, recall that the t distribution is similar to

the standard normal, so the P-value can be approximated using the normal probability table.

EXAMPLE 9.10 Regression for Selling Price of Homes

What affects the selling price of a house? Table 9.4 shows observations on home sales in Gainesville, Florida, in fall 2006. This table shows data for 8 homes. The entire file for 100 home sales is the "house selling price" data file at the text Web site. Variables listed are selling price (in dollars), size of house (in square feet), annual taxes (in dollars), number of bedrooms, number of bathrooms, and whether the house is newly built. For now, we use only the data on y = selling price and x = size of house.

TABLE 9.4: Selling Prices and Related Factors for a Sample of Home Sales in Gainesville, Florida

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: For the complete file for 100 homes, see the text Web site.

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 9.15 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 are

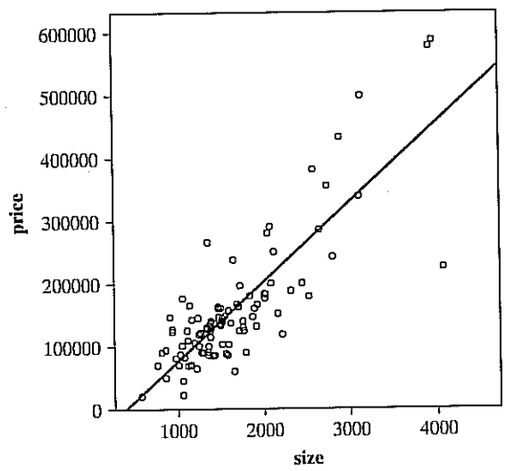

FIGURE 9.15: Scatterplot and Prediction Equation for y = Selling Price (in Dollars) and x = Size of House (in Square Feet)

regression outliers, however, and one point falls quite far below the overall trend. We discuss this abnormality in Section 14.5, which introduces an alternative model that does not assume constant variability around the regression line.

Table 9.5 shows part of a SPSS printout for a regression analysis. The prediction equation is $\hat{y} = -50,926 + 126.6x$. The predicted selling price increases by $b = 126.6$ dollars for an increase in size of a square foot. Figure 9.15 also superimposes the prediction equation over the scatterplot. In SPSS, "Beta" denotes the estimated standardized regression coefficient. For the regression model of this chapter, this is the correlation; it is not to be confused with the population slope, β, which is unknown.

TABLE 9.5: Information from SPSS Printout for Regression Analysis of y = Selling Price and x = Size of House

```
                N          Mean       Std. Deviation
-------------------------------------------------------
price         100     155331.00            101262.21
size          100       1629.28               666.94

                Sum of
                Squares      df    Mean Square
Regression    7.057E+11       1    7.057E+11
Residual      3.094E+11      98    3157352537
Total         10.15E+11      99

    R Square          Std. Error of the Estimate
        .695                  56190.324

                Unstandardized   Standardized
                  Coefficients   Coefficients
                     B    Std. Error      Beta         t        Sig.
(Constant)     -50926.3   14896.373                 -3.42       .001
size           126.594        8.468       .834      14.95       .000
```

Table 9.5 reports that the standard error of the slope estimate is $se = 8.47$. This is listed under "Std. Error" for the size predictor. 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 the last row of Table 9.5. 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 9.5 under the heading "Sig", 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.

Table 9.6 shows part of a SAS printout for the same analysis. The two-sided P-value, listed under the heading "Pr $>|t|$," is <0.0001 to four decimal places. (It is

TABLE 9.6: Part of a SAS Printout for Regression Analysis of Selling Price and Size of House

Source	DF	Sum of Squares	Mean Square
Model	1	7.05729E11	7.05729E11
Error	98	3.094205E11	3157352537
Corrected Total	99	1.01515E12	

Root MSE 56,190.3

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-50926	14896	-3.42	0.0009
size	1	126.59411	8.46752	14.95	<.0001

actually 0.0000000... to a huge number of decimal places, but SAS reports it this way rather than 0.0000 so you don't think the P-value is *exactly* 0.)

Both the SAS and SPSS printouts 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, there is extremely strong evidence 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

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 from Table B, 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.

EXAMPLE 9.11 Estimating the Slope 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 $t_{.025} = 1.984$. (Table B shows $t_{.025} = 1.984$ for $df = 100$.) 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. To obtain the confidence interval for a constant multiple of the slope (such as 100β, the change in the mean of y for an increase of 100 units in x), multiply the endpoints of the confidence interval for β by the same constant.

For Table 9.4, 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.

Reading the Computer Printout

Let's take a closer look at the printouts in Tables 9.5 and 9.6. They contain some information we have not yet discussed. For instance, in the sum of squares table, the sum of squared errors (SSE) is 3.094 times 10^{11}. This is a huge number because the y-values are very large and their deviations are squared. The estimated conditional standard deviation of y is

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

SAS labels this "Root MSE" for square root of the mean square error. SPSS misleadingly labels it "Std. Error of the Estimate." This is a poor label, because s refers to a conditional standard deviation of selling prices (for a fixed house size), not a standard error of a statistic.

The sum of squares table also reports the total sum of squares, $\text{TSS} = \sum(y - \bar{y})^2 = 10.15 \times 10^{11}$. From this value and SSE,

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

This is the proportional reduction in error in using house size to predict selling price. Since the slope of the prediction equation is positive, the correlation is the positive square root of this value, or 0.834. A strong positive association exists between these variables.

The total sum of squares TSS partitions into two parts, the sum of squared errors, $\text{SSE} = 3.094 \times 10^{11}$, and the difference between TSS and SSE, $\text{TSS} - \text{SSE} = 7.057 \times 10^{11}$. This difference is the numerator of the r^2 measure. SPSS calls this the ***regression sum of squares***. SAS calls it the ***model sum of squares***. 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.

The table of sums of squares has an associated list of degrees of freedom values. The degrees of freedom for the total sum of squares $\text{TSS} = \sum(y - \bar{y})^2$ is $n - 1 = 99$, since TSS refers to variability in the *marginal* distribution of y, which has sample

variance $s_y^2 = \text{TSS}/(n-1)$. The degrees of freedom for SSE equals $n-2=98$, since SSE refers to variability in the *conditional* distribution of y, which has variance estimate $s^2 = \text{SSE}/(n-2)$ for a model having two parameters. The regression (or 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 SSE equals $df = n-1$ for the total sum of squares, in this case $1+98=99$.

Inference for the Correlation*

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 testing H_0: $\rho=0$ is

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

This provides the same value as the test statistic $t=b/se$. Use either statistic to test H_0: independence, since each has the same t sampling distribution with $df=n-2$ and yields the same P-value. For example, the correlation of $r=0.834$ for the house selling price data has

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

This is the same t test statistic as Example 9.10 (page 278) had for testing H_0: $\beta=0$.

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 9.7 shows the way software reports the correlation matrix for the variables selling price of home, 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 P-value for testing H_0: $\rho=0$ against H_a: $\rho\neq 0$ is listed beneath the correlation.

TABLE 9.7: Correlation Matrix for House Selling Price Data. Value under correlation is two-sided P-value for testing H_0: $\rho=0$

Correlations / P-value for Ho: Rho=0

	price	size	taxes	bedrooms
price	1.00000	0.83378 <.0001	0.84198 <.0001	0.39396 <.0001
size	0.83378 <.0001	1.00000	0.81880 <.0001	0.54478 <.0001
taxes	0.84198 <.0001	0.81880 <.0001	1.00000	0.47393 <.0001
bedrooms	0.39396 <.0001	0.54478 <.0001	0.47393 <.0001	1.00000

The correlations on the diagonal running from the upper left-hand corner to the lower right-hand corner of a correlation matrix all equal 1.000. 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.

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.

Missing Data

In a correlation analysis, some subjects may not have observations for one or more of the variables. For example, Table 9.13 in the exercises lists 10 variables for 40 nations. Observations on a few of the variables, such as literacy rate, are missing for several nations.

For statistical analyses, some software deletes all subjects for which data are missing on at least one variable. This is called ***listwise deletion***. Other software only deletes a subject for analyses for which that observation is needed. For example, this approach uses a subject in finding the correlation for two variables if that subject provides observations for both variables, regardless of whether the subject provides observations for other variables. This approach is called ***pairwise deletion***. With this approach, the sample size can be larger for each analysis.

These days, more sophisticated and better strategies exist than both of these. They are not yet available in most software, and they are beyond the scope of this text. For details, see Allison (2002).

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 9.16 illustrates a straight line falling close to an actual curvilinear relationship.

The inferences discussed 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 9.5. 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 an association exists. In summary, $\beta=0$ need not correspond to independence

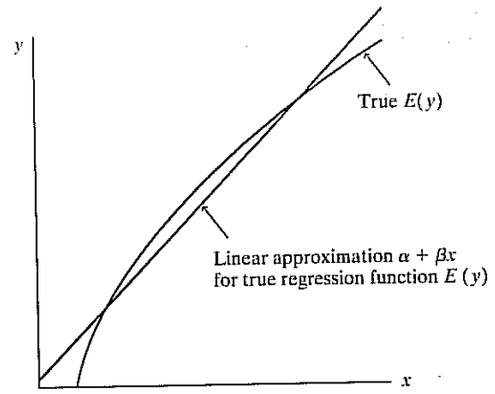

FIGURE 9.16: A Linear Regression Equation as an Approximation for Nonlinear Relationship

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 (that is, 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 making modifications to the analysis, if necessary.

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 might 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.

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 9.17 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 9.17 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.

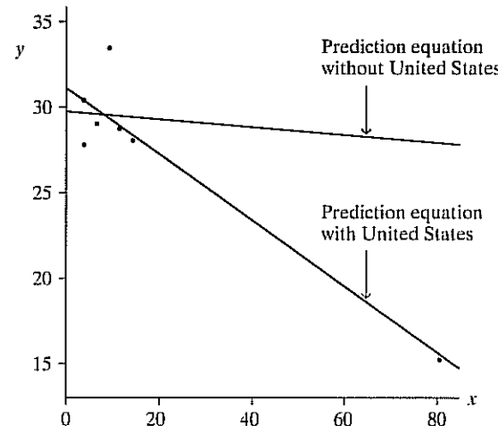

FIGURE 9.17: Prediction Equations for y = Birth Rate and x = Television Ownership, with and without Observation for United States

Section 9.2 showed a not-so-extreme version of this. The slope of the prediction equation more than doubled when we included the observation for D.C. in the data set about statewide murder rates.

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 9.17. It may suggest an additional 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, as we did following Example 9.4 (page 261) with the D.C. observation for the murder rates.

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 9.17, the correlation is -0.935 when the U.S. 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 9.18 shows a scatterplot of 500 points that is regular and has a correlation of $r = 0.71$. Suppose, instead, we had only sampled the middle half of the points, roughly between 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 9.15, $r = 0.834$. 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.

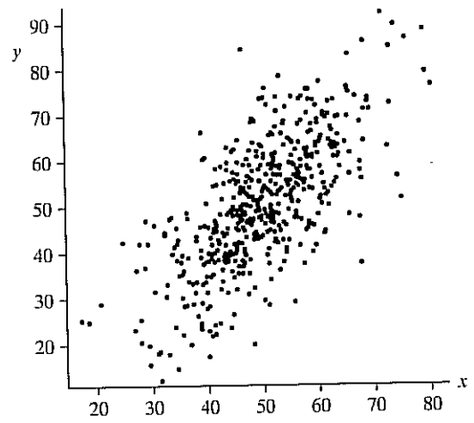

FIGURE 9.18: The Correlation is Affected by the Range of x-Values. The correlation decreases from 0.71 to 0.33 using only points with x between 43 and 57.

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 randomly sampled from the population of *all* high school students who take the SAT and placed 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 9.18, 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.

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 mean

(i.e., above the regression line) or below that mean (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 9.19. 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.

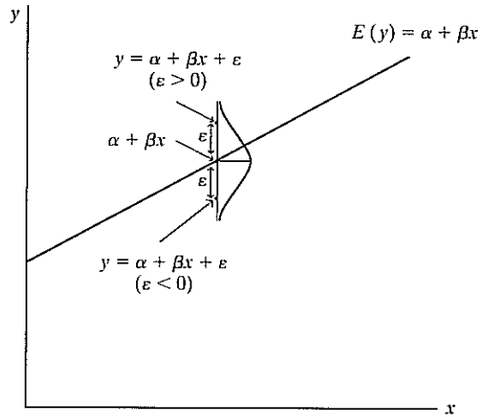

FIGURE 9.19: Positive and Negative ε-Values Correspond to Observations above and below the Mean of the Conditional Distribution

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 that $e = y - \hat{y}$. Then e is the *residual*, the difference between the observed and predicted values of y. 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 and Reality

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. By definition, 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 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 in terms of a 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 for the slope estimates the size of the effect.

Table 9.8 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 categorical response variables. Before discussing these multivariate models, however, we introduce in the next chapter some new concepts that help us to understand and interpret multivariate relationships.

TABLE 9.8: Summary of Tests of Independence and Measures of Association

	Measurement Levels of Variables		
	Nominal	Ordinal	Interval
Null hypothesis	H_0: Independence	H_0: Independence	H_0: Independence ($\beta = 0$)
Test statistic	$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$	$z = \frac{\hat{\gamma}}{se}$	$t = \frac{b}{se}, df = n - 2$
Measure of association	$\hat{\pi}_2 - \hat{\pi}_1$	$\hat{\gamma} = \frac{C-D}{C+D}$	$r = b\left(\frac{s_x}{s_y}\right)$
	Odds ratio		$r^2 = \frac{\text{TSS}-\text{SSE}}{\text{TSS}}$

PROBLEMS

Practicing the Basics

9.1. For the following variables in a regression analysis, which variable more naturally plays the role of x (explanatory variable) and which plays the role of y (response variable)?
- **(a)** College grade point average (GPA) and high school GPA
- **(b)** Number of children and mother's education level
- **(c)** Annual income and number of years of education
- **(d)** Annual income and assessed value of home

9.2. Sketch plots of the following prediction equations, for values of x between 0 and 10:
- **(a)** $\hat{y} = 7 + 0.5x$
- **(b)** $\hat{y} = 7 + x$
- **(c)** $\hat{y} = 7 - x$
- **(d)** $\hat{y} = 7 - 0.5x$
- **(e)** $\hat{y} = 7$
- **(f)** $\hat{y} = x$

9.3. Anthropologists often try to reconstruct information using partial human remains at burial sites. For instance, after finding a femur (thighbone), they may want to predict how tall an individual was. An equation they use to do this is $\hat{y} = 61.4 + 2.4x$, where $\hat{y}$ is the predicted height and x is the length of the femur, both in centimeters.[3]
- **(a)** Identify the y-intercept and slope of the equation. Interpret the slope.
- **(b)** A femur found at a particular site has length of 50 cm. What is the predicted height of the person who had that femur?

9.4. The OECD (Organization for Economic Cooperation and Development) consists of 20 advanced, industrialized countries. For these nations,[4] the prediction equation relating y = child poverty rate in 2000 to x = social expenditure as a percent of gross domestic product is $\hat{y} = 22 - 1.3x$. The y-values ranged from 2.8% (Finland) to 21.9% (U.S.). The x-values ranged from 2% (U.S.) to 16% (Denmark).
- **(a)** Interpret the y-intercept and the slope.
- **(b)** Find the predicted poverty rates for the U.S. and for Denmark.
- **(c)** The correlation is -0.79. Interpret.

9.5. Look at Figure 2 in www.ajph.org/cgi/reprint/93/4/652?ck=nck, a scatterplot for U.S. states with correlation 0.53 between x = child poverty rate and y = child mortality rate. Approximate the y-intercept and slope of the prediction equation shown there.

9.6. A study[5] of mail survey response rate patterns of the elderly found a prediction equation relating x = age (between about 60 and 90) and y = percentage of subjects responding of $\hat{y} = 90.2 - 0.6x$.
- **(a)** Interpret the slope.
- **(b)** Find the predicted response rate for a (i) 60-year-old, (ii) 90-year-old.

9.7. For recent UN data from 39 countries on y = per capita carbon dioxide emissions (metric tons per

[3] S. Junger, *Vanity Fair*, October 1999.

[4] *Source*: Figure 8H in www.stateofworkingamerica.org

[5] D. Kaldenberg et al., *Public Opinion Quarterly*, Vol. 58, 1994, p. 68.