Chapter 9

Linear Regression and Correlation

Table 9.1 shows recent data from *Statistical Abstract of the United States* for the 50 states and the District of Columbia on several variables:

- 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 the population living in metropolitan areas.
- Percentage of the population who are white.
- Percentage of the population who are high school graduates or higher.
- Percentage of families living below the poverty level.
- Percentage of families headed by a single parent (male householders with no wife present and with own children, or female householders with no husband present and with own children).

All these variables are quantitative. Crime rate and murder rate are natural response variables. One might treat the other variables as explanatory variables for these responses.

This chapter introduces methods for analyzing relationships between a pair of quantitative variables We present three different, but related, aspects of such relationships:

TABLE 9.1 Statewide Data Used to Illustrate Regression Analyses

State	Violent Crime Rate	Murder Rate	Metropolitan Residents	Percent White	High School Graduates	Poverty Rate	Single Parent
AK	761	9.0	41.8	75.2	86.6	9.1	14.3
AL	780	11.6	67.4	73.5	66.9	17.4	11.5
AR	593	10.2	44.7	82.9	66.3	20.0	10.7
AZ	715	8.6	84.7	88.6	78.7	15.4	12.1
CA	1078	13.1	96.7	79.3	76.2	18.2	12.5
CO	567	5.8	81.8	92.5	84.4	9.9	12.1
CT	456	6.3	95.7	89.0	79.2	8.5	10.1
DE	686	5.0	82.7	79.4	77.5	10.2	11.4
FL	1206	8.9	93.0	83.5	74.4	17.8	10.6
GA	723	11.4	67.7	70.8	70.9	13.5	13.0
HI	261	3.8	74.7	40.9	80.1	8.0	9.1
IA	326	2.3	43.8	96.6	80.1	10.3	9.0
ID	282	2.9	30.0	96.7	79.7	13.1	9.5
IL	960	11.4	84.0	81.0	76.2	13.6	11.5
IN	489	7.5	71.6	90.6	75.6	12.2	10.8
KS	496	6.4	54.6	90.9	81.3	13.1	9.9
KY	463	6.6	48.5	91.8	64.6	20.4	10.6
LA	1062	20.3	75.0	66.7	68.3	26.4	14.9
MA	805	3.9	96.2	91.1	80.0	10.7	10.9
MD	998	12.7	92.8	68.9	78.4	9.7	12.0
ME	126	1.6	35.7	98.5	78.8	10.7	10.6
MI	792	9.8	82.7	83.1	76.8	15.4	13.0
MN	327	3.4	69.3	94.0	82.4	11.6	9.9
MO	744	11.3	68.3	87.6	73.9	16.1	10.9
MS	434	13.5	30.7	63.3	64.3	24.7	14.7
MT	178	3.0	24.0	92.6	81.0	14.9	10.8
NC	679	11.3	66.3	75.2	70.0	14.4	11.1
ND	82	1.7	41.6	94.2	76.7	11.2	8.4
NE	339	3.9	50.6	94.3	81.8	10.3	9.4
NH	138	2.0	59.4	98.0	82.2	9.9	9.2
NJ	627	5.3	100.0	80.8	76.7	10.9	9.6
NM	930	8.0	56.0	87.1	75.1	17.4	13.8
NV	875	10.4	84.8	86.7	78.8	9.8	12.4
NY	1074	13.3	91.7	77.2	74.8	16.4	12.7
OH	504	6.0	81.3	87.5	75.7	13.0	11.4
OK	635	8.4	60.1	82.5	74.6	19.9	11.1
OR	503	4.6	70.0	93.2	81.5	11.8	11.3
PA	418	6.8	84.8	88.7	74.7	13.2	9.6
RI	402	3.9	93.6	92.6	72.0	11.2	10.8
SC	1023	10.3	69.8	68.6	68.3	18.7	12.3
SD	208	3.4	32.6	90.2	77.1	14.2	9.4
TN	766	10.2	67.7	82.8	67.1	19.6	11.2
TX	762	11.9	83.9	85.1	72.1	17.4	11.8
UT	301	3.1	77.5	94.8	85.1	10.7	10.0
VA	372	8.3	77.5	77.1	75.2	9.7	10.3
VT	114	3.6	27.0	98.4	80.8	10.0	11.0
WA	515	5.2	83.0	89.4	83.8	12.1	11.7
WI	264	4.4	68.1	92.1	78.6	12.6	10.4
WV	208	6.9	41.8	96.3	66.0	22.2	9.4
WY	286	3.4	29.7	95.9	83.0	13.3	10.8
DC	2922	78.5	100.0	31.8	73.1	26.4	22.1

1. We investigate *whether an association exists* between the two variables by testing the hypothesis of statistical independence.
2. We study the *strength of their association* using a measure of association called the *correlation*.
3. We study the *form of the relationship*. Using the data, we estimate a formula that predicts a subject's score on the response variable from the score on the explanatory variable. For instance, a formula of this type predicts a state's murder rate from the percentage of its population living below the poverty level.

The analyses conducted in studying these three aspects of the relationship between two quantitative variables are collectively called a ***regression analysis***. Section 9.1 introduces the use of a straight line to describe the form of the relationship. Section 9.2 presents the method, called *least squares*, that estimates the best line for a particular data set. Section 9.3 describes the *linear regression model*, which takes into account variability of the data about the straight line. Section 9.4 defines the *Pearson correlation* for describing the strength of a linear relationship. Section 9.5 presents 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

The univariate methods of Chapters 5 and 6 summarized quantitative variables by their means. As noted in Chapters 7 and 8, most analyses involve more than one variable and distinguish between response and explanatory variables. This chapter presents methods for analyzing a pair of quantitative variables, one of which is a response variable and one of which is an explanatory variable.

Notation for Response and Explanatory Variables

Let Y denote the *response* variable and let X denote the *explanatory* variable.

We shall analyze how the response variable Y tends 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 type of formula describes how Y = murder rate varies among states for different levels of X = percent below the poverty level. For instance, does the murder rate tend to be higher for states that have higher poverty levels?

Linear Functions

A variety of different formulas might describe how Y relates to X. Any particular formula might provide a good description of the relationship or a poor one. This chapter

introduces the simplest class of such formulas, *straight lines*. They are called ***linear functions***.

Linear Function

The formula $Y = \alpha + \beta X$ expresses the response variable Y as a ***linear function*** of the explanatory variable X. The formula maps out 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. This 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, the value $X = 0$ has $Y = 3 + 2(0) = 3$; the value $X = 1$ has $Y = 3 + 2(1) = 5$. Figure 9.1 plots this function. The horizontal axis, called the ***X-axis***, lists the possible values of the explanatory variable, X; the vertical axis, called the ***Y-axis***, lists the possible values of Y. The axes intersect at the point where $X = 0$ and $Y = 0$, called the *origin*. □

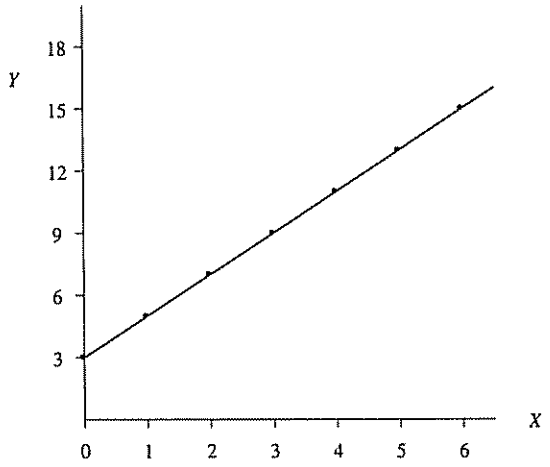

Figure 9.1 Graph of the Straight Line $Y = 3 + 2X$. The Y-intercept is 3 and the slope is 2.

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 $Y = 3 + 2X$, at $X = 0$ and $X = 1$, the Y values are 3 and 5. These Y values differ by $5 - 3 = 2$, which is the value of β for this line. Similarly, 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 of a straight line.

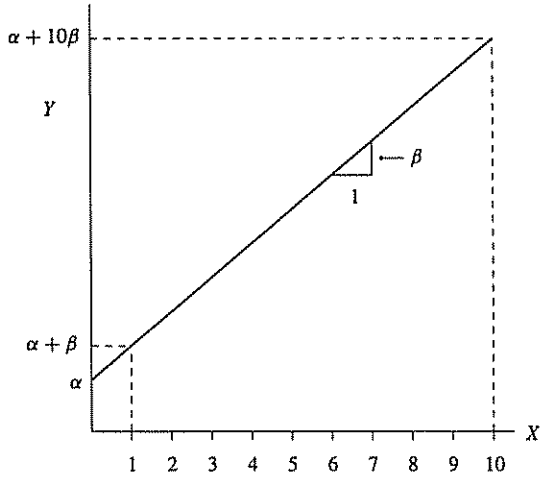

Figure 9.2 Graph of the Straight Line $Y = \alpha + \beta X$. The Y-intercept is α and the slope is β.

One can draw the straight line by finding any two separate pairs of (X, Y) values on the graph and then drawing the line through the points. For instance, we could 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 $Y = 3 + 2X$.

The slope describes the rate of change of the Y-values as X increases. A one-unit increase in X corresponds to a change of β units in Y, no matter what the value of X. The larger the absolute value of β, the steeper the line. A line with $\beta = 4$, such as $Y = 3 + 4X$, climbs more quickly than one with $\beta = 2$.

Example 9.2 Straight Line for Violent Crime and Poverty

For the 50 states, consider the variables Y = violent crime rate (number of violent crimes per 100,000 population) and X = poverty rate (percentage of state residents living below the poverty level). The straight line $Y = 210 + 25X$ approximates the relation between these variables. 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 poverty rate increases by one percent, the violent crime rate

increases by about 25 (i.e., 25 crimes per 100,000 population). Similarly, if the poverty rate increases by 10%, the crime rate increases by about 10(25) = 250.

By contrast, if instead the explanatory variable X refers to the percentage of the population living in metropolitan 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. A 1% change in poverty rate has a greater effect on the violent crime rate than a 1% change in metropolitan residence. Figure 9.3 shows the lines relating the violent crime rate to poverty rate and metropolitan residence. □

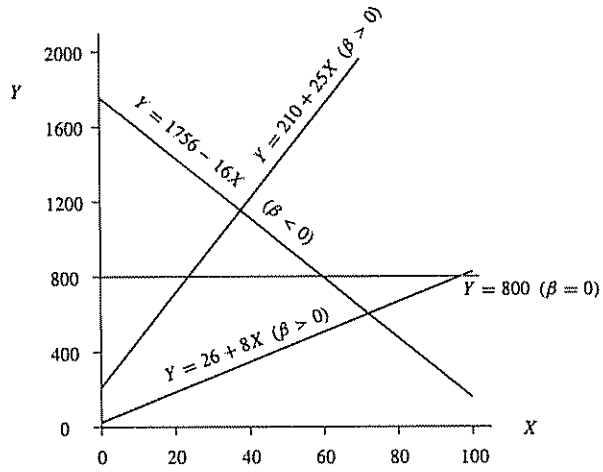

Figure 9.3 Graphs of Lines Showing Positive Relationships ($\beta > 0$), a Negative Relationship ($\beta < 0$), and Independence ($\beta = 0$)

If β is positive, then Y increases as X increases; that is, large values of Y occur with large values of X, and small values of Y occur with small values of X. The straight line then goes upward, like the lines just mentioned relating violent crime rate to poverty rate or metropolitan residence. 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 $Y = 1756 - 16X$ approximates the relationship between Y = violent crime rate and X = percentage of residents who are high school graduates. This shows a negative relationship with slope -16. 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 of a linear function 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, then 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

The linear function is the simplest mathematical function and provides the simplest form for the relationship between two quantitative variables. The function $Y = \alpha + \beta X$ is a ***model***. That is, the formula provides a simple approximation for the true relationship between X and Y. For a given value of X, the model predicts a value for Y. The better these predictions tend to be, the better the model.

9.2 Least Squares Prediction Equation

Using sample data, one can estimate the linear model relating Y and X. The process treats α and β in the linear function $Y = \alpha + \beta X$ as unknown parameters and yields estimates of these parameters. The estimated linear function then provides predictions about Y at fixed values for X.

Scatter Diagram

The first step of model fitting is to look at the data. A plot of the data reveals whether a model with a straight line trend makes sense.

The values (X, Y) of the two variables for any particular subject form a point relative to the X and Y axes. To portray graphically the sample relationship, we plot the sample of n observations on X and Y as n points. This graphical plot is called a ***scatter diagram*** or ***scatterplot***.

Example 9.3 Scatter Diagram for Murder Rate and Poverty

We use the data from Table 9.1 on X = poverty rate and Y = murder rate throughout the chapter to illustrate aspects of regression analysis. To check whether Y is approximately linearly related to X, we first construct a scatter diagram for the 51 observations. Figure 9.4 shows this plot.

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. We plot the point (9.7, 12.7) by moving from the origin 9.7 units in the X (horizontal) direction and 12.7 units in the Y (vertical) direction. This brings us to the spot 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 D.C., for which the murder rate was much higher than that of 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 the results of a regression analysis. □

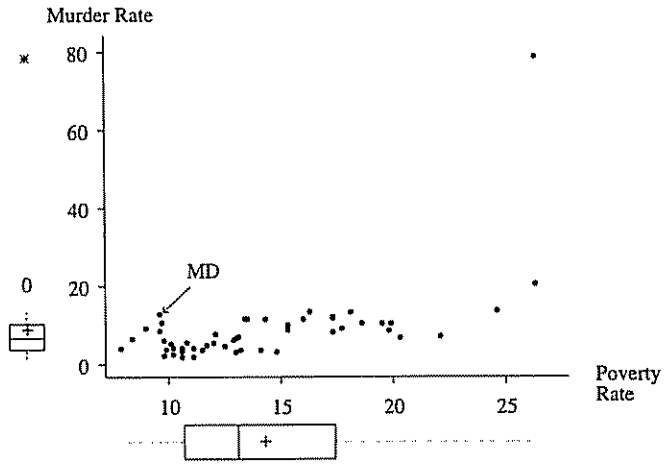

Figure 9.4 Scatter Diagram for Y = Murder Rate and X = Percentage of Residents Below the Poverty Level, for 50 States and D.C.

The scatter diagram provides a visual means of checking whether a relationship is approximately linear. When the relationship seems strongly nonlinear, it does not make sense to use a straight line to model the data. Figure 9.5 illustrates such a case. This figure shows a negative relationship over part of the range of X values, and a positive relationship over the rest. These cancel each other out using a linear model. For such data, one needs a different type of model, presented in Section 14.4.

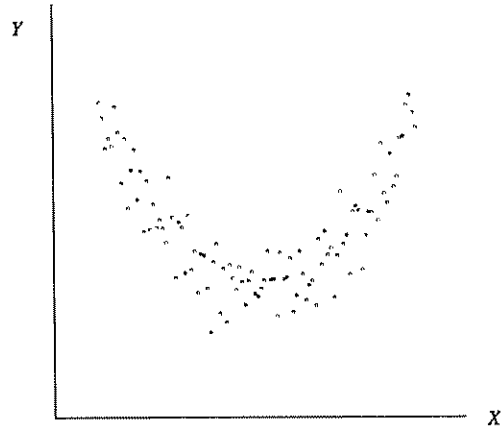

Figure 9.5 A Nonlinear Relationship, for Which It Is Inappropriate to Use a Straight Line Regression Model

Prediction Equation

When the scatter diagram suggests that the linear model $Y = \alpha + \beta X$ is realistic, we estimate this unknown line. The notation

$$\hat{Y} = a + bX$$

represents a *sample* equation, based on the data, that estimates the linear function $Y = \alpha + \beta X$. In the sample equation, the Y-intercept (a) estimates the unknown Y-intercept α of the linear model, and the slope (b) estimates the unknown slope β.

Substituting a particular value of X into the formula $a + bX$ provides a value, denoted by $\hat{Y}$, that predicts the value of Y at that value of X. The sample equation $\hat{Y} = a + bX$ is called the ***prediction equation***, since it provides predictions about the response variable Y for any value of X.

The prediction equation is the best straight line summarizing the trend of the points in the scatter diagram. It falls closest to those points, in a certain average sense discussed later in this section. The formulas for the coefficients of this prediction equation 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.

We shall not dwell on these formulas or even illustrate how to use them. The formula for b is complex and is messy to use even for a small to moderate sample size. Although a somewhat simpler computational formula exists, anyone who does any serious regression modeling these days uses a computer or a calculator that has these formulas programmed. For the extensions of this linear model studied in later chapters, it is impractical to calculate prediction equations even with a calculator but very simple to do so using statistical software. We strongly encourage that you use modern computing methods, rather than hand calculation, to conduct regression analyses. To use statistical software on a PC or other type of computer, you supply the data set and either select the desired methods from a menu or enter certain code specifying those methods (e.g., regression) and the options desired (e.g., scatter diagrams, prediction equation). The appendix at the end of the text provides details.

Example 9.4 Prediction Equation for Murder Rate and Poverty

Using statistical software for the 51 observations on murder rate and poverty rate from Table 9.1, we get the results shown in the sample printout in Table 9.2. Usually, software provides summary statistics such as the mean and standard deviation for each variable, as well as statistics for the regression analysis. For instance, murder rate has $\bar{Y} = 8.7$ and $s = 10.7$, indicating that it is probably highly skewed to the right. In fact, the box plot for murder rate in Figure 9.4 shows that the extreme outlying observation for D.C. contributes a lot to this outcome.

The estimates of α and β are listed under the heading "Parameter Estimate." The estimate of the Y-intercept is $a = -10.14$, listed opposite "INTERCEP." The estimate

TABLE 9.2 Part of Computer Printout for Fitting Linear Regression Model to 51 Observations (Including D.C.) on X = Percent in Poverty and Y = Murder Rate

Variable	Mean	Std Dev
MURDER	8.727	10.718
POVERTY	14.259	4.584

Dependent Variable: MURDER

Variable	Parameter Estimate	Standard Error
INTERCEP	-10.1364	4.1206
POVERTY	1.3230	0.2754

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 relating Y = murder rate to X = poverty rate is $\hat{Y} = a + bX = -10.14 + 1.32X$.

For this prediction equation, the slope is $b = 1.32$. Since the slope is positive, the sample relationship is positive. The larger the poverty rate, the larger the murder rate tends to be. The value of 1.32 for the slope indicates that, on the average, an increase of 1% in the poverty rate relates to an increase of 1.32 in the murder rate.

Similarly, an increase of 10% in the poverty rate corresponds to a $10(1.32) = 13.2$-unit increase in the murder rate. If one state has a 12% poverty rate and another has a 22% poverty rate, for example, the predicted 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 as if 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 Prediction Equation

Figure 9.6 plots the prediction equation from Example 9.4 over the scatter diagram. The diagram shows that one observation, the one for D.C., is a serious outlier. This outlier appears to have a substantial effect on the results, since the line seems to be pulled up toward it and away from the center of the general trend of points.

Table 9.3 shows part of a computer printout from using the observations for the 50 states but not the one for D.C. The prediction equation equals $\hat{Y} = -.86 + .58X$. Figure 9.6 also shows this line, which passes more directly through the 50 points. The slope is .58, compared to 1.32 when the observation for D.C. is included. The one outlying observation has the effect of more than doubling the slope! An observation is called ***influential*** if removing it results in a large change in the prediction equation. Unless the sample size is large, an observation can have a strong influence on the slope if its X value is low or high compared to the rest of the data.

In summary, the line for the data set including the D.C. observation seems to distort the relationship for the other 50 states. It seems wiser to use the equation based on data

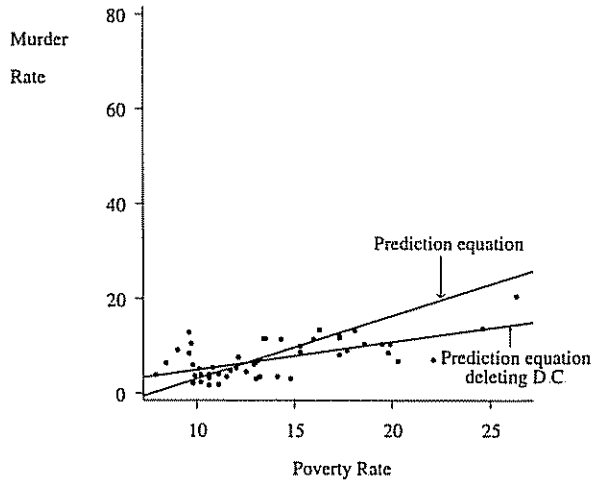

Figure 9.6 Prediction Equations Relating Murder Rate and Percentage in Poverty, with and without D.C. Observation

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 for the states. 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.

Residuals

The prediction equation $\hat{Y} = -.86 + .58X$ provides predicted murder rates for states with various levels of X = poverty rate. For the sample data, a comparison of the predicted values to the actual murder rates checks the goodness of the prediction equation.

For example, consider Massachusetts, for which $X = 10.7$ and $Y = 3.9$. The predicted murder rate ($\hat{Y}$) at $X = 10.7$ is $\hat{Y} = -.86 + .58X = -.86 + .58(10.7) = 5.4$. The prediction error is the difference between the actual Y value of 3.9 and the predicted value of 5.4, or $Y - \hat{Y} = 3.9 - 5.4 = -1.5$. Thus, the prediction equation overestimates the murder rate by 1.5. Similarly, for Louisiana, $X = 26.4$ and $\hat{Y} = -.86 + .58(26.4) = 14.6$. 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***. For a given observation, the response variable equals Y and the predicted value equals $\hat{Y} = a + bX$. The difference $Y - \hat{Y}$ is the residual for that observation.

Residual

The difference between observed and predicted values of the response variable, $Y - \hat{Y}$, is called a ***residual***.

TABLE 9.3 Part of Computer Printout for Fitting Linear Model to 50 States (Excluding D.C.) on X = Percent in Poverty and Y = Murder Rate

Source	DF	Sum of Squares	Mean Square	Variable	Parameter Estimate
Model	1	307.342	307.34	INTERCEP	-0.8567
Error	48	470.406	9.80	POVERTY	0.5842
Total	49	777.749			

Root MSE 3.1305

Obs	Dep Var MURDER	Predict Value	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
5	13.1000	9.7765	3.3235
6	5.8000	4.9273	0.8727

Table 9.3 shows the murder rates, the predicted values, and the residuals for some of the states. A positive residual results when the observed value Y is larger than the predicted value $\hat{Y}$, and $Y - \hat{Y} > 0$. 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.

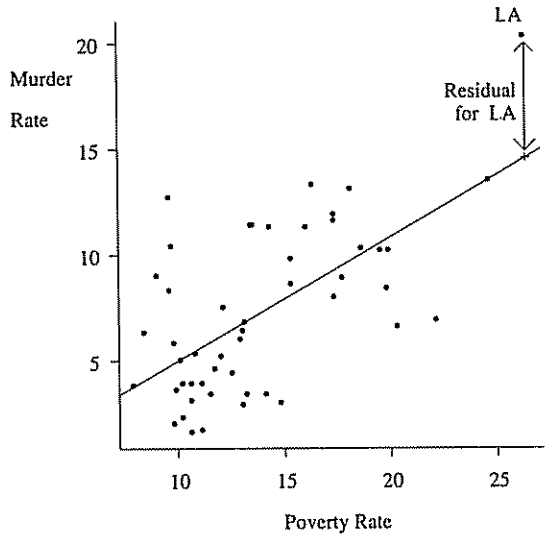

Figure 9.7 Prediction Equation and Residuals

Graphically in the scatter diagram, the residual for an observation is the vertical distance between the point for the observation and the prediction line. Figure 9.7 illustrates this for the murder rate data for the 50 states. 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 corresponding to substituting $X = 26.4$ into the prediction equation $\hat{Y} = -.86 + .58X$. The residual is the difference between the observed and predicted points, which is the vertical distance $Y - \hat{Y} = 20.3 - 14.6 = 5.7$.

Method of Least Squares

Each observation in a sample has a residual. Some residuals are positive, some are negative, and their average equals 0. Just as deviations of observations from their mean $\bar{Y}$ satisfy $\sum(Y - \bar{Y}) = 0$, so is the prediction equation defined so that $\sum(Y - \hat{Y}) = 0$.

If the prediction line falls close to the points in the scatter diagram, the residuals tend to be 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 the ***sum of squared errors***. This terminology refers to the use of the residual as a measure of prediction error from using $\hat{Y}$ to predict Y. The measure SSE is also called the ***residual sum of squares***. It describes the variation of the observed points around the prediction line.

The better the prediction equation, the smaller the residuals tend to be and, hence, the smaller the summary measure SSE tends to be. Any particular equation has a corresponding set of residuals and a value of SSE. The prediction equation specified by the usual formulas for estimates a and b of α and β is the one with the smallest value of SSE out of all possible linear prediction equations.

Method of Least Squares, Least Squares Estimates

The ***method of least squares*** provides the prediction equation $\hat{Y} = a + bX$ having the minimal value of $\text{SSE} = \sum(Y - \hat{Y})^2$. The ***least squares estimates*** a and b are the values determining the prediction equation for which the sum of squared errors SSE is a minimum.

The best prediction equation, according to the least squares criterion, is the one with the smallest sum of squared residuals. Because of this criterion, the prediction line $\hat{Y} = a + bX$ is called the ***least squares line***. If we square the residuals (such as those in Table 9.3) for the least squares line $\hat{Y} = -.86 + .58X$ and then sum them, we get

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

This value of SSE is smaller than the value of SSE for *any* other straight line prec such as $\hat{Y} = -.88 + .60X$. In this sense, the observed points fall closer to this lin to *any* other line. Software for regression lists the value of SSE. Table 9.3 report the "Sum of Squares" column, in the row labeled "Error." In some software, su SPSS, this is labeled as "Residual" in the sum of squares column.

9.3 The Linear Regression Model

For the model $Y = \alpha + \beta X$, each value of X corresponds to a single value of Y. a model is said to be ***deterministic***. It is unrealistic in social science research, *variability* occurs in Y-values among subjects with the same X-value.

For example, let X = number of years of education and Y = annual income. sider the subjects having $X = 12$ years of education. They do not all have the income, since income is not completely dependent upon education. Instead, th a probability distribution of annual income for individuals with $X = 12$. This bution refers to the variability in the Y values at a *fixed* value of X, so it is a c ***tional distribution***. Similarly, a separate conditional distribution applies for thos $X = 13$ years of education, and others apply for those with each possible value The conditional distributions of income differ at the different levels of educatio mean of the conditional distribution probably is higher at higher levels of educa

A ***probabilistic*** model for the relationship between X and Y is one that allo variability in the values of Y at each value of X. We now show how a linear fu is the basis for a probabilistic model.

Linear Regression Function

A probabilistic model uses $\alpha + \beta X$ to represent the *mean* of Y, rather than Y itse function of X. For a given value of X, $\alpha + \beta X$ represents the mean of the cond 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 repres *expected value*, which is another term for the *mean*.

We now use the form of equation

$$E(Y) = \alpha + \beta X$$

to model the relationship between X and the mean of the conditional distributio For instance, let Y = violent crime rate and X = poverty rate for the 50 state probabilistic model $E(Y) = 210 + 25X$ is more plausible than the deterministic $Y = 210 + 25X$. At $X = 10$, for instance, $E(Y) = 210 + 25(10) = 460$. For having 10% of their residents below the poverty level, the model predicts that th

violent crime rate is 460. Different states having $X = 10$ would have different violent crime rates, but the average response would be about 460.

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, since it uses a straight line for the relationship between the mean of Y and 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

For each fixed value of X, there is a conditional distribution of Y-values. The mean of that distribution equals $E(Y) = \alpha + \beta X$. The linear regression model has an additional parameter σ describing the standard deviation of that 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***.

Example 9.5 Income Variation for Education Predictor

Suppose the linear regression model $E(Y) = -5 + 3X$ with $\sigma = 13$ describes the relationship between Y = annual income, in thousands of dollars, and X = number of years of education. According to this model, for individuals with X years of education, their incomes have a mean of $E(Y) = -5 + 3X$ and a standard deviation of 13.

For instance, those having a high school education ($X = 12$) have a mean income of $E(Y) = -5 + 3(12) = 31$ and a standard deviation of 13. If the conditional distribution of income is approximately bell-shaped, then about 95% of the incomes fall between \$5000 and \$57,000. These are the values that are two standard deviations from the mean; that is, $31 - 2(13) = 5$ and $31 + 2(13) = 57$. Similarly, those with a college education ($X = 16$) have a mean annual income of $E(Y) = -5 + 3(16) = 43$ thousand dollars, with about 95% of the incomes falling between \$17,000 and \$69,000.

The slope of $\beta = 3$ implies that mean income increases 3 thousand dollars for each year increase in education. Figure 9.8 pictures this linear regression model with $\alpha = -5$, $\beta = 3$, and $\sigma = 13$. That figure plots the regression function $E(Y) = -5 + 3X$ and the conditional income distributions at $X = 8$, 12, and 16 years. □

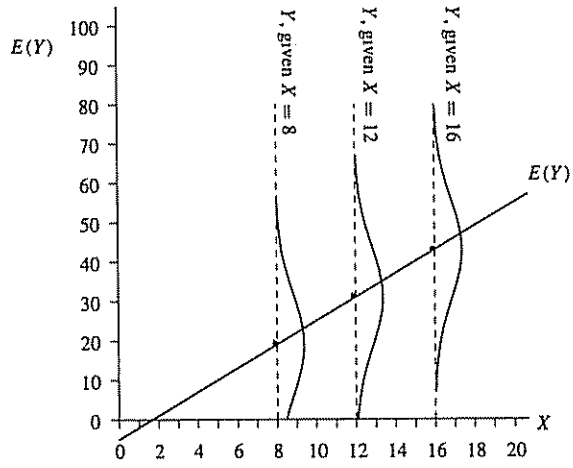

Figure 9.8 The Regression Model $E(Y) = -5 + 3X$, with $\sigma = 13$, Relating Y = Income (in Thousands of Dollars) to X = Education (in Years)

In Figure 9.8, each conditional distribution is normal, and each has the same conditional standard deviation, $\sigma = 13$. In practice, of course, the conditional distributions would not be exactly normal, and the standard deviation need not be the same for each. In fact, *any particular model never holds exactly in practice*. It is merely a simple approximation for reality. When we have sample data, we can check whether a particular model is realistic by checking the fit of the prediction equation on the scatter diagram.

Estimating Conditional Variation

This section has assumed that the mean of Y is related to X by a *linear* regression equation and that it, therefore, makes sense to use a straight line to estimate that equation. The optimality of the least squares estimates is based partly on the further assumption that the standard deviation σ of the conditional distribution of Y is identical at the various values of X.

The conditional standard deviation σ refers to the variability of Y values about the conditional population mean $E(Y) = \alpha + \beta X$ for all subjects with the same value of X. To estimate σ, we use the numerical value for $\text{SSE} = \sum(Y - \hat{Y})^2$, which measures sample variability about the least squares line. The estimate is

$$\hat{\sigma} = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n-2}}$$

Example 9.6 Estimating Variability for Murder Rates

We return to the analysis of Y = murder rate and X = poverty rate, for the 50 states. From Section 9.2 and Table 9.3, the prediction equation is $\hat{Y} = -.86 + .58X$, and the sum of squared errors in using X to predict Y equals $\text{SSE} = 470.4$. The estimate of the conditional standard deviation is

$$\hat{\sigma} = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{470.4}{50-2}} = 3.1$$

The model predicts that at any fixed value X of poverty rate, the murder rates vary around a mean of $-.86 + .58X$ with a standard deviation of 3.1.

At $X = 20$, for instance, the conditional distribution of murder rates has an estimated mean of $-.86 + .58(20) = 10.7$ and standard deviation of 3.1. For an approximately bell-shaped distribution, about 95% of the observations fall within two standard deviations of the mean. Then, for a particular value of X, most murder rates fall within about $2\hat{\sigma} = 2(3.1) = 6.2$ of the true mean murder rate at that value of X. □

Chapter 14 shows how to check the constant variation assumption, through analysis of the residuals. A quick check with the scatter diagram of whether the variability around the fitted line is fairly constant for various values of X is also a way to evaluate whether the assumption is reasonable. If it is, $\hat{\sigma}$ describes the variability of the observations about the regression line, at any fixed X value. Otherwise, $\hat{\sigma}$ provides a measure of *average* variability about the line.

Conditional Versus Marginal Variation

From Sections 3.3 and 5.1, a sample point estimate of the standard deviation of a variable Y is given by

$$s = \sqrt{\frac{\sum(Y - \bar{Y})^2}{n - 1}}$$

This is the standard deviation of the *marginal* distribution of Y, since the formula makes no reference to any other variable X. To emphasize that the standard deviation of the marginal distribution of Y 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, which we denote by $\hat{\sigma}$ for the sample and σ for the population.

For the data on murder rates for the 50 states, the marginal standard deviation is $s_Y = 4.0$ (It is 10.7 if D.C. is included!), whereas the conditional standard deviation is $\hat{\sigma} = 3.1$. Typically, there is less spread in Y at a fixed value of X than there is totaled over all such values. Thus, the conditional standard deviation of Y is usually less than the marginal standard deviation of Y.

For another example, the *marginal* distribution of college GPAs (Y) at a particular university may primarily fall between about 1.0 and 4.0, perhaps having a standard deviation of $\sigma_Y = .60$. For those students with a high school GPA (X) of 3.8, however, the *conditional* distribution of college GPAs might cluster between 2.7 and 4.0, perhaps having $\sigma = .25$. Figure 9.9 illustrates the distinction between the marginal and conditional distributions and their standard deviations, the conditional distribution typically being less disperse.

The term $(n - 2)$ in the denominator of $\hat{\sigma}$ is called the ***degrees of freedom*** (df) for the estimate. In general, when a regression equation has p unknown parameters, then $df = n - p$. The linear model $E(Y) = \alpha + \beta X$ has two parameters (α and β), so $df = n - 2$. The problem studied in Chapter 5 of estimating a single mean μ corresponds to estimating the parameter in the simpler regression model, $E(Y) = \mu$,

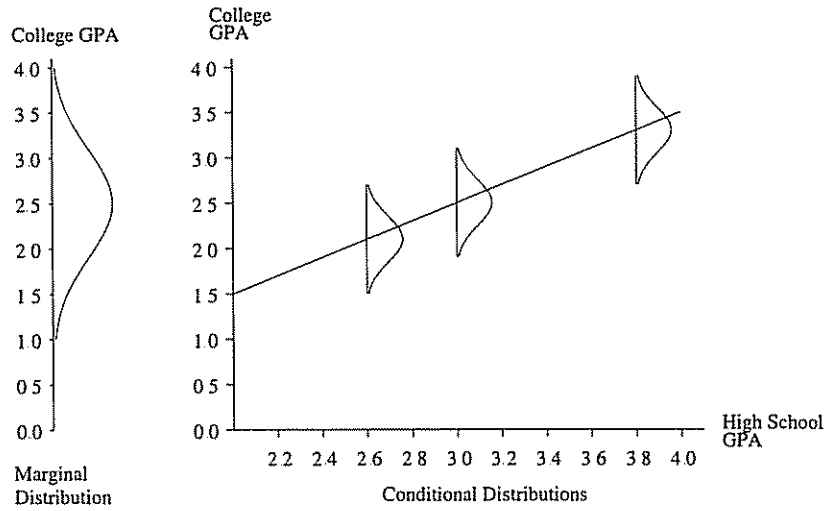

Figure 9.9 Marginal and Conditional Distributions

with a single parameter. Thus, the estimate s_Y of the standard deviation of the marginal distribution has $df = n - 1$.

Table 9.3 lists SSE = 470.4 and its degrees of freedom, $df = n - 2 = 50 - 2 = 48$. The ratio of these, $\hat{\sigma}^2 = 9.80$, is listed on the printout and called the "Mean Square Error." This is often abbreviated by MSE. Its square root is the estimate of the conditional standard deviation of Y, $\hat{\sigma} = \sqrt{9.80} = 3.13$, which is listed under the heading "ROOT MSE."

9.4 Measuring Linear Association—The Correlation

The linear regression model uses a straight-line prediction equation to describe the relationship between two quantitative variables. This section introduces two measures of the strength of linear 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, it indicates 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 the numerical value of the slope is intrinsically linked to the units of measurement of the variables.

For example, consider the prediction equation $\hat{Y} = -.86 + .58X$ for Y = murder rate and X = poverty rate for the 50 states. Suppose we measure murder rate as the number of murders per 1,000,000 population instead of 100,000 population. A one-

unit increase in the poverty rate corresponds to a $b = .58$ increase in the predicted number of murders per 100,000 people, which is equivalent to a 5.8 increase in the predicted number of murders per 1,000,000 people. Thus, if Y = number of murders per 1,000,000 population, the slope of the prediction equation is 5.8 instead of .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 the response variable differed.

In summary, the slope b of the prediction equation depends on the units of measurement. It doesn't directly indicate whether the association is strong or weak, since 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 for the 50 states relating the murder rate to percentage living in metropolitan areas is $3.28 + .06X$. A one-unit increase in the percentage living in metropolitan areas corresponds to a .06 increase in the murder rate, whereas a one-unit increase in the percentage below the poverty level corresponds to a .58 increase in the murder rate. A change of 1% in poverty rate has a much greater effect on murder rate than a change of 1% in metropolitan residence.

The measures of association we now study do not vary according to the units of measurement. Like the measures of association presented in the previous chapter for categorical data, their magnitudes indicate the strength of association.

The Pearson Correlation

The measure of association for quantitative variables known as the ***Pearson correlation***, or simply the ***correlation***, is a *standardized* version of the slope. It is a type of slope for which the value, unlike that of b, does not depend on the units of measurement. The standardization adjusts the slope b for the fact that the marginal distributions of X and Y have standard deviations that depend on the units of measurement for X and Y. The correlation is the value the slope assumes if the measurement units for the two variables are such that their standard deviations are equal.

Let s_X and s_Y denote the sample standard deviations of X and Y,

$$s_X = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}} \quad \text{and} \quad s_Y = \sqrt{\frac{\sum(Y - \bar{Y})^2}{n-1}}$$

Pearson Correlation

The ***Pearson correlation***, denoted by r, is related to the slope b of the prediction equation $\hat{Y} = a + bX$ by

$$r = \left(\frac{s_X}{s_Y}\right) b$$

Multiplying the slope by the ratio of the sample standard deviations provides a standardized measure. When the sample dispersions are equal ($s_X = s_Y$), the correlation is

the same as the slope. The most common case of this is when the variables are standardized by converting their values to z-scores, in which case both standardized variables have standard deviations of 1.0. Because of this relationship between r and b, the Pearson correlation is also called the ***standardized regression coefficient*** for the regression model $E(Y) = \alpha + \beta X$.

Example 9.7 Correlation Between Murder Rate and Poverty Rate

For the 50 states, the prediction equation relating Y = murder rate to X = poverty rate is $\hat{Y} = -.86 + .58X$. From ordinary calculations for a standard deviation, or better yet, reading the value off a printout, the standard deviation of poverty rate is $s_X = 4.29$, and the standard deviation of murder rate is $s_Y = 3.98$. The Pearson correlation for these data equals

$$r = \left(\frac{s_X}{s_Y}\right) b = \left(\frac{4.29}{3.98}\right)(.58) = .63$$

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

Properties of the Pearson Correlation

- The correlation is valid only when a straight line is a reasonable 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 standardized version of the slope, unlike b, must fall between -1 and $+1$. The reason for this will be seen later in the section.
- r has the same sign as b. Since r equals the slope 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. Also, $r = 0$ for those lines having $b = 0$. So, when $r = 0$, there is not a linear increasing or linear decreasing trend in the relationship between X and Y.
- $r = \pm 1$ 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 the value of Y.
- The larger the absolute value of r, the stronger the degree of linear association. Two variables with a correlation of $-.80$ are more strongly linearly associated than two variables with a correlation of .40, for example. Figure 9.10 shows scatter diagrams having a variety of values for r. Again, we emphasize *linear* association, since r is not designed to describe the relationship when it is curvilinear. In that case, as shown in Figure 9.11, the least squares line may be completely or nearly horizontal, and $r = 0$ when $b = 0$. A low absolute value for r does not then imply that the variables are unassociated, but that the association is not linear.
- 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 $r = .63$.

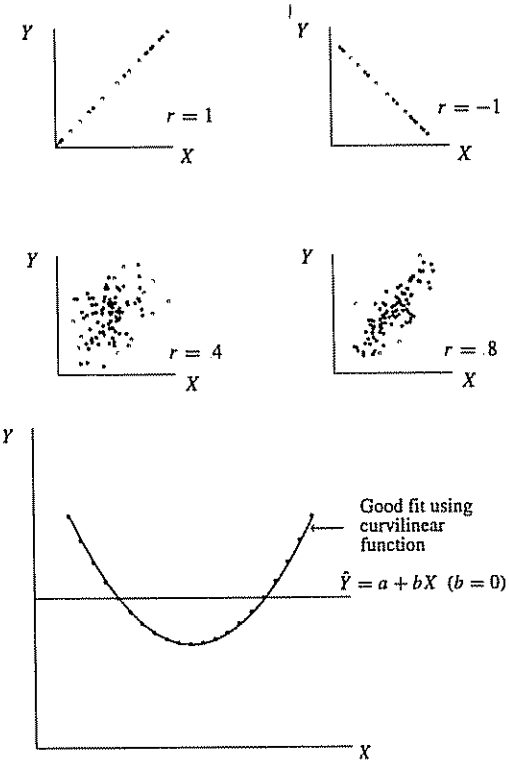

Figure 9.10 Scatter Diagrams for Different Correlation Coefficients

Figure 9.11 Scatter Diagram for Which $r = 0$, Even Though There Is a Strong Curvilinear Relationship

- 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. When murder rate predicts poverty rate, the correlation is the same as when poverty rate predicts murder rate; $r = .63$ in both cases.
- We now further interpret the correlation as a standardized slope. We can rewrite the equality $r = (s_X/s_Y)b$ as $s_X b = rs_Y$. Now, the slope b is the change in the predicted value of Y for a one-unit increase in X. An increase in X of s_X units relates to a change in the predicted value of Y 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 one standard deviation in X corresponds to a change of r standard deviations in the Y variable. The larger the absolute value of r, the stronger the association, in the sense that a standard deviation change in X corresponds to a greater proportion of a standard deviation change in Y.

Example 9.8 Regression Toward the Mean

Sir Francis Galton, the British statistical scientist who discovered the basic ideas of regression in the 1880s, noted that the correlation between X = father's height and Y = son's height (or between X = mother's height and Y = daughter's height) is about

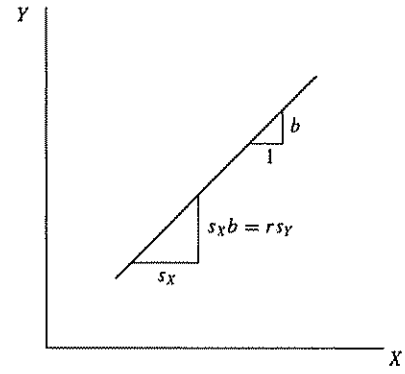

Figure 9.12 An Increase of s_X Units in X Corresponds to a Change of rs_Y Units in Y

.5. From the last property, a standard deviation change in parent's height corresponds to half a standard deviation change in child's height.

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

Since r is less than 1, a Y value is predicted to be fewer standard deviations from its mean than X is from its mean. Tall parents tend to have tall children, but on the average not quite so tall. For instance, if you consider all fathers with height 7 feet, perhaps their sons average 6 feet 5 inches—taller than average, but not so extremely tall; if you consider all fathers with height 5 feet, perhaps their sons average 5 feet 5 inches—shorter than average, but not so extremely short. In each case, there is a *regression toward the mean*. This is the origin of the name for regression analysis. □

For the data on X = poverty rate and Y = murder rate, the correlation of $r = .63$ implies that the variables are positively related. A standard deviation increase in the poverty rate corresponds to a .63 standard deviation increase in murder rate. The change of .63 standard deviations is independent of whether murder rate is measured as murders per 1000 population, murders per 10,000 population, or whatever. It is the same *real* amount regardless of the units of measurement. If r had been only .20, say, the association would have been much weaker; a standard deviation increase in poverty rate then corresponds to less than one-third as large a change in the predicted murder rate.

The correlation is useful for comparing relationships for variables measured with 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. Thus, their slopes from the separate prediction equations are not comparable. The correlations are comparable. Suppose the correlation of education with murder rate is $-.30$. Since the correlation of poverty rate with murder rate is .63,

and since $.63 > |-.30|$, murder rate is more strongly associated with poverty rate than with education.

The properties of the Pearson correlation are similar to those of the ordinal measure of association *gamma* (Section 8.5). It falls between -1 and $+1$, and larger absolute values relate to stronger associations. The correlation of .63 between murder rate and poverty rate is moderately strong. Incidentally, the Pearson of *Pearson correlation* is Karl Pearson, the British statistical scientist who proposed r as an estimate of the true correlation in 1896, just four years before proposing his chi-squared test of independence for contingency tables.

A Proportional Reduction in Error Measure

Another measure of association between two quantitative variables uses the proportional reduction in error (PRE) formulation introduced in Section 8.7. The strength of association between an explanatory variable X and a response variable Y is judged by the goodness of X as a predictor of Y. If one 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 strongly related.

This measure of association, like all PRE measures, 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 number of errors with the two rules is $E_1 - E_2$. Converting this reduction in error to a proportion provides the definition of the measure as the proportional reduction in error,

$$\text{PRE} = \frac{E_1 - E_2}{E_1}$$

Rule 1 (Predicting Y without using X): The best predictor is $\bar{Y}$, the sample mean. The one number that is closest to all the observations on Y, in an aggregate sense, is $\bar{Y}$.

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. If X is correlated with Y, this provides better predictions than the crude predictor $\bar{Y}$ for rule 1, which ignores the tendency for Y to change as X changes.

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 called the ***total sum of squares*** of the Y-values about their mean. We denote thi by TSS. For rule 2, the predicted values are the $\hat{Y}$ values from the prediction equatio The total prediction error equals

$$E_2 = \sum \left(Y - \hat{Y}\right)^2$$

We previously denoted this by SSE, the ***sum of squared errors***.

When a strong linear relationship exists between X and Y, the prediction equa tion provides predictions ($\hat{Y}$) that are much better than $\bar{Y}$, in the sense that the sum c squared prediction errors is substantially less. Figure 9.13 shows graphical representa tions of the two predictors and their prediction errors. For rule 1, the same predictio ($\bar{Y}$) applies for the value of Y, regardless of the value of X. For rule 2 the predictio changes as X changes, and the prediction errors tend to be smaller.

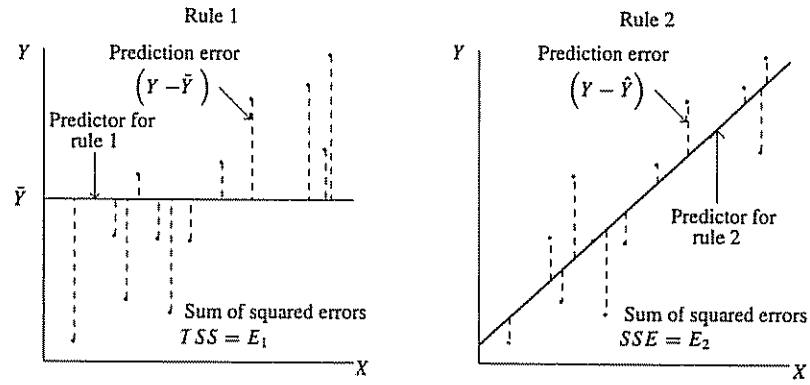

Figure 9.13 Graphical Representation of rule 1 and Total Sum of Squares $E_1 = TSS = \sum(Y - \bar{Y})^2$, rule 2 and Residual Sum of Squares $E_2 = SSE = \sum(Y - \hat{Y})^2$

Definition of Measure: The proportional reduction in error from using the linear prediction equation instead of $\bar{Y}$ to predict Y is called the ***coefficient of determination***. It is denoted by r^2. The PRE formula for r^2 is

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

Properties of Coefficient of Determination

We use the notation r^2 for this measure because, in fact, it equals the square of the Pearson correlation r. If we know the correlation, it is simple to calculate the coefficient of determination. The PRE formula above is useful for interpreting r^2, but it is not needed for its calculation.

The coefficient of determination, like the Pearson correlation, measures the strength of *linear* association between X and Y. We emphasize *linear* because r^2 compares the linear prediction equation to $\bar{Y}$ for making predictions.

Example 9.9 r^2 for Murder and Poverty Rates

The correlation between poverty rate and murder rate for the 50 states is $r = .629$. The coefficient of determination is, therefore, $r^2 = (.629)^2 = .395$. For predicting murder rate, the linear prediction equation $\hat{Y} = -.86 + .58X$ has 39.5% less error than $\bar{Y}$.

One can also calculate r^2 directly from the definition. Software for regression routinely provides tables of sums of squares, such as shown in part of Table 9.3. From it, the sum of squared errors using the prediction equation is $\text{SSE} = \sum(Y - \hat{Y})^2 = 470.4$; the total sum of squares is $\text{TSS} = \sum(Y - \bar{Y})^2 = 777.7$ (Alternatively, if we already know $s_Y = 3.98$, then $\sum(Y - \bar{Y})^2 = (n-1)s_Y^2 = 49(3.98)^2$). Thus,

$$r^2 = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = \frac{777.7 - 470.4}{777.7} = \frac{307.3}{777.7} = .395$$

Normally, it is unnecessary to perform this computation, since most software reports r or r^2 or both. □

The properties of the coefficient of determination r^2 follow directly from those of the Pearson correlation r.

- Since $-1 \le r \le 1$, r^2 falls between 0 and 1.
- The minimum possible value for SSE is 0, in which case $r^2 = \text{TSS}/\text{TSS} = 1$. For 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 $r = \pm 1$.
- If SSE = TSS, then $r^2 = 0$, in which case the slope $b = 0$ also. Now, SSE can be no larger than TSS, since the least squares line provides the minimum sum of squared errors about a line. When the least squares slope $b = 0$, the Y-intercept a equals $\bar{Y}$ (since $a = \bar{Y} - b\bar{X}$, which equals $\bar{Y}$ when $b = 0$), so that $\hat{Y} = \bar{Y}$ for all X. The two prediction rules are then identical, so that SSE = TSS and $r^2 = 0$.
- 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 relative to $\bar{Y}$ in predicting the response variable.
- r^2 does not depend on the units of measurement.
- r^2 takes the same value when X predicts Y as when Y predicts X.

Conditional Versus Marginal Variability

To summarize, the Pearson 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, indicating what the slope equals for units of measurement such that X and Y are equally disperse. A one standard deviation change in X corresponds to an r standard deviation change in Y. The square of the correlation has a proportional reduction in error interpretation related to the reduction in error from predicting Y using $\hat{Y} = a + bX$ rather than $\bar{Y}$.

The total sum of squares, $\text{TSS} = \sum(Y - \bar{Y})^2$, summarizes the *variability* of the observations on Y, since this quantity divided by $n - 1$ is the sample variance s_Y^2 of the Y-values. Similarly, $\text{SSE} = \sum(Y - \hat{Y})^2$ summarizes the variability around the prediction equation, which refers to variability for the conditional distributions. When $r^2 = .39$, 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 "X explains 39% of the variability in Y" or "39% of the variance in Y is explained by its linear relationship with X." Roughly speaking, the variance of the conditional distribution of Y for a given X is 39% smaller than the variance of the marginal distribution of Y.

When two variables X and Y are strongly associated, the variation in the conditional distributions is considerably less than the variation in the marginal distribution. For instance, the estimated conditional standard deviation $\hat{\sigma}$ is then much smaller than the sample marginal standard deviation s_Y of Y. Figure 9.9 illustrated this.

9.5 Inferences for the Slope and Correlation

Sections 9.1–9.3 showed how a linear function can represent the *form* of relationships between quantitative variables. Section 9.4 used the Pearson 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 parameters in regression models.

A test of whether the two quantitative variables are statistically independent has the same purpose as the chi-squared test for qualitative variables. More usefully, one can construct a confidence interval for the slope β of the regression equation or for the population Pearson correlation. 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

The statistical inferences presented in this section make the following assumptions:

- The mean of Y is related to X by the linear equation $E(Y) = \alpha + \beta X$.
- The conditional standard deviation is identical at each X-value.
- The conditional distribution of Y at each value of X is normal.
- The sample is selected randomly.

The first assumption states that the linear regression model is valid. The second assumption is one under which the least squares estimates are optimal. The third assumption is needed for the standardized version of the sample slope or correlation to have a t sampling distribution. In practice, of course, none of these assumptions is ever satisfied exactly. For large samples, the normality assumption is unimportant. If the second assumption is violated, other estimates may be more efficient than least squares, but the methods we discuss here are still approximately valid. The first and last 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.

Test of Independence

Under these assumptions, suppose the 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 model $E(Y) = \alpha + \beta X$, this means that the slope β of the regression line equals 0 (see Figure 9.14). The null hypothesis that the variables are statistically independent is $H_0 : \beta = 0$.

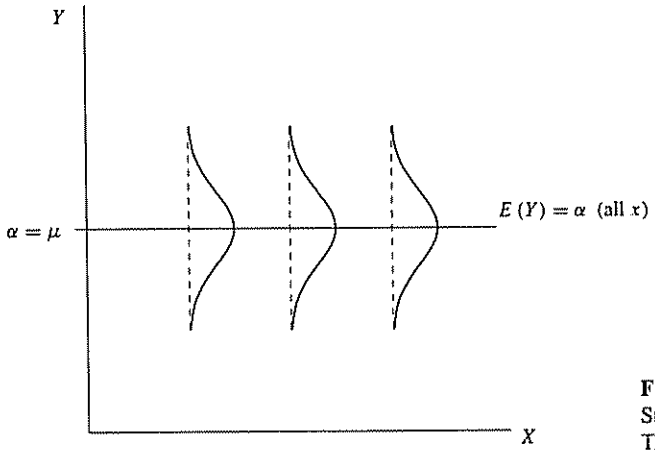

Figure 9.14 X and Y Are Statistically Independent When the True Slope Equals 0

We normally test independence against the two-sided alternative hypothesis H_a: $\beta \neq 0$. Occasionally we use a one-sided alternative, H_a: $\beta > 0$ or H_a: $\beta < 0$, to predict the direction of the association. The test statistic in either case equals

$$t = \frac{b}{\hat{\sigma}_b}$$

where $\hat{\sigma}_b$ denotes 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 of the parameter ($\beta = 0$), 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 formula for the estimated standard error of b is

$$\hat{\sigma}_b = \frac{\hat{\sigma}}{\sqrt{\sum(X - \bar{X})^2}} = \frac{\hat{\sigma}}{s_X\sqrt{n-1}}$$

This depends on the point estimate $\hat{\sigma}$ of the standard deviation of the conditional distributions of Y. From Section 9.3, this estimate is

$$\hat{\sigma} = \sqrt{\frac{\text{SSE}}{n-2}}$$

called the ***root mean square error*** by most software.

The smaller the standard deviation of the conditional distributions, the more precisely b estimates β. A small $\hat{\sigma}$ occurs when the data points show little variability about the prediction equation. Also, the standard error of b is inversely related to $\sum(X-\bar{X})^2$, the sum of squares of the observed X-values about their mean. This sum increases, and hence b estimates β more precisely, as the sample size n increases and when the X-values are more highly spread out.

As in the two-sided t test for a mean, the P-value for the alternative $H_a : \beta \neq 0$ is the two-tail probability from the t distribution (Table B). The degrees of freedom for the t sampling distribution are the same as the df of the estimate $\hat{\sigma}$, namely, the denominator $n - 2$. When $df > 30$, recall that the t distribution is very similar to the standard normal distribution. We can then treat the t test statistic as a z statistic and approximate the P-value using the normal probability table (Table A). Computer software provides the appropriate P-value for any value of df.

Example 9.10 Regression for Selling Price of Homes

What factors affect the selling price of a house? Table 9.4 shows 93 observations on home sales in Gainesville, Florida, in January 1996. Variables listed are selling price (in thousands of dollars), size of house (in thousands of square feet), 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.

Since these 93 observations come from one city alone, we cannot use them to make inferences about the relationship between X and Y in general. The houses sold are not even a random sample of homes in Gainesville. Nevertheless, 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 scatter diagram of selling price and size of house, with a box plot for each variable. The diagram displays a strong positive trend, and the model $E(Y) = \alpha + \beta X$ seems appropriate. A couple of the points are severe outliers on each variable, however, and one of them falls rather far from the overall trend; we discuss this abnormality in Section 14.5, which introduces an alternative model for these data that does not assume constant variability around the regression line.

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

P	*S*	*Be*	*Ba*	*New*	*P*	*S*	*Be*	*Ba*	*New*	*P*	*S*	*Be*	*Ba*	*New*
48.5	1.10	3	1	0	76.0	1.66	3	2	0	108.0	1.79	4	2	1
55.0	1.01	3	2	0	81.8	1.33	3	2	0	107.5	1.85	3	2	0
68.0	1.45	3	2	0	84.5	1.34	3	2	0	109.9	2.06	4	2	1
137.0	2.40	3	3	0	83.5	1.40	3	2	0	110.0	1.76	4	2	0
309.4	3.30	4	3	1	86.0	1.15	2	2	1	120.0	1.62	3	2	1
17.5	.40	1	1	0	86.9	1.58	3	2	1	115.0	1.80	4	2	1
19.6	1.28	3	1	0	86.9	1.58	3	2	1	113.4	1.98	3	2	0
24.5	.74	3	1	0	86.9	1.58	3	2	1	114.9	1.57	3	2	0
34.8	.78	2	1	0	87.9	1.71	3	2	0	115.0	2.19	3	2	0
32.0	.97	3	1	0	88.1	2.10	3	2	0	115.0	2.07	4	2	0
28.0	.84	3	1	0	85.9	1.27	3	2	0	117.9	1.99	4	2	0
49.9	1.08	2	2	0	89.5	1.34	3	2	0	110.0	1.55	3	2	0
59.9	.99	2	1	0	87.4	1.25	3	2	0	115.0	1.67	3	2	0
61.5	1.01	3	2	0	87.9	1.68	3	2	0	124.0	2.40	4	2	0
60.0	1.34	3	2	0	88.0	1.55	3	2	0	129.9	1.79	4	2	1
65.9	1.22	3	1	0	90.0	1.55	3	2	0	124.0	1.89	3	2	0
67.9	1.28	3	2	0	96.0	1.36	3	2	1	128.0	1.88	3	2	1
68.9	1.29	3	2	0	99.9	1.51	3	2	1	132.4	2.00	4	2	1
69.9	1.52	3	2	0	95.5	1.54	3	2	1	139.3	2.05	4	2	1
70.5	1.25	3	2	0	98.5	1.51	3	2	0	139.3	2.00	4	2	1
72.9	1.28	3	2	0	100.1	1.85	3	2	0	139.7	2.03	3	2	1
72.5	1.28	3	1	0	99.9	1.62	4	2	1	142.0	2.12	3	3	0
72.0	1.36	3	2	0	101.9	1.40	3	2	1	141.3	2.08	4	2	1
71.0	1.20	3	2	0	101.9	1.92	4	2	0	147.5	2.19	4	2	0
76.0	1.46	3	2	0	102.3	1.42	3	2	1	142.5	2.40	4	2	0
72.9	1.56	4	2	0	110.8	1.56	3	2	1	148.0	2.40	5	2	0
73.0	1.22	3	2	0	105.0	1.43	3	2	1	149.0	3.05	4	2	0
70.0	1.40	2	2	0	97.9	2.00	3	2	0	150.0	2.04	3	3	0
76.0	1.15	2	2	0	106.3	1.45	3	2	1	172.9	2.25	4	2	1
69.0	1.74	3	2	0	106.5	1.65	3	2	0	190.0	2.57	4	3	1
75.5	1.62	3	2	0	116.0	1.72	4	2	1	280.0	3.85	4	3	0

Note: P = selling price (thousands of dollars), S = size (thousands of square feet), Be = number of bedrooms, Ba = number of bathrooms, $New = 1$ if new and 0 if not. Data provided by Jane Myers, Coldwell-Banker Realty.

Table 9.5 shows part of a SAS computer printout for a regression analysis. The least squares estimates of α and β are $a = -25.2$ and $b = 75.6$, so the prediction equation is $\hat{Y} = -25.2 + 75.6X$. The predicted selling price increases by $b = 75.6$ thousand dollars for an increase in size of 1 thousand square feet; that is, it increases about \$76 per square foot. Figure 9.15 also superimposes the prediction equation over the scatter diagram.

Table 9.5 reports that the standard error of the slope estimate is $\hat{\sigma}_b = 3.865$. This is listed under "Standard Error." This value estimates the variability in sample slope values that would result from repeatedly selecting random samples of 93 house sales in Gainesville and calculating prediction equations.

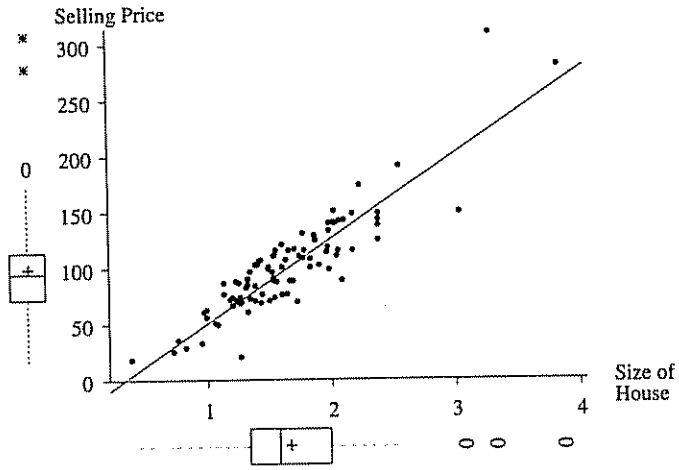

Figure 9.15 Scatter Diagram and Prediction Equation for Y = Selling Price of House (in Thousands of Dollars) and X = Size of House (in Thousands of Square Feet)

To test independence, $H_0 : \beta = 0$, we form the t test statistic,

$$t = \frac{b}{\hat{\sigma}_b} = \frac{75.61}{3.865} = 19.6$$

This statistic is listed in Table 9.5 under "T for H0: Parameter = 0." Since the sample size equals $n = 93$, the degrees of freedom for this statistic are $df = n - 2 = 91$. This t statistic has sampling distribution that is essentially identical to the standard normal distribution. Clearly, this is an extremely large test statistic.

The P-value, listed in Table 9.5 under the heading PROB > $|T|$, is .0001 to four decimal places. This refers to the two-sided alternative $H_a : \beta \neq 0$; that is, 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, $|t| = 19.6$, if H_0 were true. For the one-sided alternative $H_a : \beta > 0$, this level is halved, since P is then the right-hand tail probability of $t > 19.6$.

Table 9.6 shows part of a SPSS printout for the same analysis. For SPSS, "B" denotes the estimated regression coefficient, and "(Constant)" is the label for the Y-intercept. "SE B" denotes the standard error of b, "Sig T" denotes the two-sided P-value for the t test for a coefficient, and "Beta" denotes the estimated standardized regression coefficient (this is the Pearson correlation for the regression model of this chapter; it is not to be confused with the population slope, β, which is unknown).

Both the SAS and SPSS printouts also contain a standard error and t test for the Y-intercept. This information is usually not of interest, since rarely is there any reason to test the hypothesis that a Y-intercept equals 0. In fact, for this example, the Y-intercept does not have any interpretation, since houses of size 0 do not exist.

TABLE 9.5 Part of a SAS Computer Printout for Regression Analysis of Selling Price and Size of House

Variable	N	Mean	Std Dev
PRICE	93	99.533	44.184
SIZE	93	1.650	0.525

Source	DF	Sum of Squares	Mean Square	Root MSE	R-square
Model	1	145097.464	145097.464	19.473	0.8079
Error	91	34508.402	379.213		
Total	92	179605.867			

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-25.1936	6.6884	-3.767	0.0003
SIZE	1	75.6068	3.8652	19.561	0.0001

TABLE 9.6 Part of a SPSS Computer Printout for Regression Analysis of Selling Price and Size of House

	DF	Sum of Squares	Mean Square
Regression	1	145097.464	145097.464
Residual	91	34508.402	379.213

R Square	.8079
Standard Error	19.4734

Variable	B	SE B	Beta	T	Sig T
SIZE	75.6068	3.8652	.8988	19.561	.0000
(Constant)	-25.1936	6.6884		-3.767	.0003

In summary, we can feel highly confident in concluding that size has a positive effect on price. On the average, selling price increases as size of house increases. This is no big surprise. Indeed, we would be shocked if these variables were independent, and a test of independence is irrelevant, for practical purposes, for these data. As is usually the case, estimation is more informative than significance testing. □

Confidence Interval for the Slope

A small P-value for $H_0 : \beta = 0$ in a test of independence suggests that the regression line has a nonzero slope. However, we should be more concerned with the size of the

slope than in knowing merely that it is not 0. If the absolute value of the slope is small, in relation to the units of measurement for X and Y, the association could be statistically significant but practically unimportant. It is more informative, therefore, to construct a confidence interval for the true slope β.

A confidence interval for the slope has the formula

$$b \pm t\hat{\sigma}_b$$

The t-score is the value from Table B, with $df = n - 2$, for the desired confidence coefficient. As usual, use the t-score with half the error probability in each tail, for instance $t_{.025}$ for a 95% confidence interval. This interval is similar in form to the confidence interval for a mean (Section 6.5). We take the estimate b and add and subtract some t multiple of the standard error of the estimate.

Example 9.11 Estimating the Slope for Housing Data

For the data on X = size of house and Y = selling price, $b = 75.6$ and $\hat{\sigma}_b = 3.87$. The parameter β refers to the change in the mean selling price for a thousand square foot increase in size (i.e., a one-unit increase in X). For a 95% confidence interval, we use the $t_{.025}$ value for $df = n - 2 = 91$, which is $t_{.025} = 1.99$. (It is permissible to use the z-score if your software does not supply t scores or confidence intervals.) The interval is

$$\begin{aligned} b \pm t_{.025}\hat{\sigma}_b &= 75.6 + 1.99(3.87) \\ &= 75.6 \pm 7.7 \quad \text{or} \quad (67.9, 83.3) \end{aligned}$$

We can be 95% confident that β lies between 67.9 and 83.3. The mean selling price increases by between \$67,900 and \$83,300 for a thousand-square-foot increase in house size; equivalently, this is \$67.90 to \$83.30 per square foot. □

A confidence interval for β may not be meaningful if a one-unit increase in X is relatively small or large, in practical terms. One should make inferences about the change in $E(Y)$ for an increase in X that is a more relevant portion of the actual range of X values. To obtain endpoints for the confidence interval for a constant multiple of the slope (e.g., $.1\beta$, the change in the mean of Y for an increase of .1 unit in X), multiply the endpoints of the interval for β by the same constant.

For Table 9.4, X = size of house has $\bar{X} = 1.65$ and $s_X = .53$. A change of 1.0 thousand square feet in size is very large. Let us estimate the effect of a 100-square-foot increase in area. Now, 100 square feet equals .1 thousand square feet, so we refer to the effect of a .1 unit change in X. The change in the mean of Y is $.1\beta$.

A confidence interval for $.1\beta$ has endpoints that are .1 times the endpoints of the interval for β. Thus, a 95% confidence interval for $.1\beta$ has endpoints $.1(67.9) = 6.79$ and $.1(83.3) = 8.33$. We infer that the mean selling price increases by at least \$6790 and at most \$8330, for a 100-square-foot increase in house size.

To illustrate, we compare the selling prices of homes having $X = 1.7$ and $X = 1.6$ thousand square feet of size. The estimated difference in selling price is $.1b =$

7.56 thousand dollars. Assuming that the linear regression model is appropriate, we conclude that the mean is between \$6790 and \$8330 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 SAS computer printout in Table 9.5. It contains a considerable amount of additional information we have not yet discussed. For instance, the middle part of the printout reports a sum of squares table. From it, the sum of squared errors (SSE) is 34,508.4. The estimated conditional standard deviation of Y for fixed X is

$$\hat{\sigma} = \sqrt{\text{SSE}/(n-2)} = \sqrt{34{,}508.4/91} = 19.5$$

This is labeled "Root MSE," for square root of the mean square error. It describes the estimated variability in selling prices, for any fixed value for size of house.

The sum of squares table also reports the total sum of squares, $\text{TSS} = \sum(Y-\bar{Y})^2 = 179,605.9$. From this value and SSE,

$$r^2 = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = \frac{179{,}605.9 - 34{,}508.4}{179{,}605.9} = .808$$

This is the proportional reduction in error in using house size to predict selling price, and is listed on the printout under the heading R-SQUARE. Since the slope of the prediction equation is positive, the Pearson correlation is the positive square root of this value, or .899. A strong positive association exists between these variables.

In Table 9.5, the total sum of squares TSS partitions into two parts, the sum of squared errors, SSE = 34,508.4, and the difference between TSS and SSE, TSS − SSE = 145,097.5. This difference is the numerator of the r^2 measure and is called the ***model sum of squares*** or sometimes the ***regression sum of squares*** or ***explained 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 the model sum of squares to the total sum of squares 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 $n - 1 = 92$, 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 $n - 2 = 91$, since it refers to variability in the *conditional* distribution of Y, which has variance estimate $\hat{\sigma}^2 = \text{SSE}/(n-2)$ for a model having two parameters. The 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 model sum of squares and *df* for the sum of squared errors equals $df = n - 1$ for the total sum of squares.

In the SPSS printout of Table 9.6, the "Regression" sum of squares is the model sum of squares and the "Residual" sum of squares is SSE. The term labeled "Standard Error" is the root mean square error, the square root of SSE divided by its *df*.

Inference for the Pearson Correlation*

The Pearson correlation r equals 0 in the same situations in which the slope b of the least squares line equals 0. Let ρ (rho) denote the value of the Pearson correlation for the regression equation $E(Y) = \alpha + \beta X$; that is, ρ is the population value of r. Then, $\rho = 0$ precisely when $\beta = 0$. Thus, 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 $t = b/\hat{\sigma}_b$. Use either statistic to test the null hypothesis of independence, since each has the same t sampling distribution with $df = n - 2$ and yields the same P-value. For example, the Pearson correlation of .899 for the house price data in Example 9.10 leads to

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = \frac{.899}{\sqrt{(1 - .808)/91}} = 19.6$$

This is the same t-value as in Example 9.10 for testing $H_0 : \beta = 0$.

For a set of variables, software for regression analyses reports their correlations in a ***correlation matrix***. This matrix is a square table listing the variables as the rows and again as the columns. It reports the correlation for each pair, and usually also the two-sided P-value for testing the significance of the correlation.

Table 9.7 illustrates the way software reports the correlation matrix for four of the variables from Table 9.1, deleting the observation for D.C. The correlation between each pair of variables appears twice. For instance, the correlation of .3688 between violent crime rate and percentage in poverty occurs both in the row for "VIOLENT" and column for "POVERTY" and in the row for "POVERTY" and column for "VIOLENT." The P-value for testing that that true correlation equals zero against the two-sided alternative is .0084, listed underneath the correlation. The correlations on the diagonal running from the upper left-hand corner to the lower right-hand corner of the table all equal 1.000, indicating simply 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.

Although tests are simple, constructing a confidence interval is more complicated for ρ 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, then the sample r cannot fall much above ρ, but it can fall well below ρ, and the sampling distribution of r is skewed to the left. Problem 9.46 shows how to construct confidence intervals for correlations.

TABLE 9.7 Computer Printout of Correlation Matrix for Variables from Table 9.1

Pearson Correlations / Prob > |R| under Ho: Rho=0

	VIOLENT	MURDER	POVERTY	SINGLE
VIOLENT	1.00000	0.78147	0.36875	0.64868
	0.0	0.0001	0.0084	0.0001
MURDER	0.78147	1.00000	0.62862	0.72807
	0.0001	0.0	0.0001	0.0001
POVERTY	0.36875	0.62862	1.00000	0.43031
	0.0084	0.0001	0.0	0.0018
SINGLE	0.64868	0.72807	0.43031	1.00000
	0.0001	0.0001	0.0018	0.0

9.6 Model Assumptions and Violations

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

Comments About Assumptions

The linear regression model assumes that the relationship between X and the mean of Y follows a straight line. The actual form is unknown, and it is unlikely to be *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. If the assumption of linearity is badly violated, as with a U-shaped relationship, results and conclusions using the linear model may be very misleading. For this reason, you should always construct a scatter diagram to check this fundamental assumption.

For instance, the inferences discussed in the previous section are appropriate for detecting positive or negative linear associations. If the true relationship were U-shaped (such as in Figure 9.5), the variables would be statistically dependent, since the mean of Y would change according to the value of X. The t test of independence might not detect it, though, since the slope b of the least squares line would be close to 0. In other words, a small P-value would probably not occur for H_0: $\beta = 0$ even though an association exists. In summary, $\beta = 0$ need not correspond to independence if the assumption of a linear regression model is violated.

The least squares line and the measures of association 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 make the additional

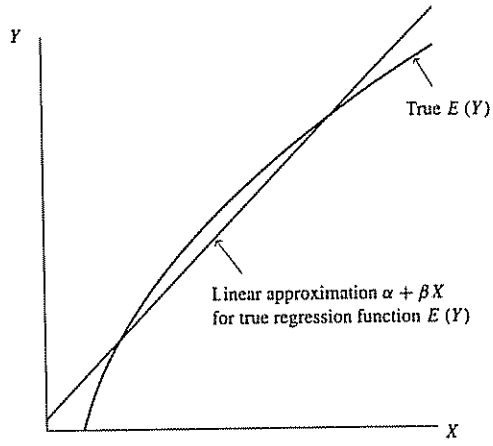

Figure 9.16 A Linear Regression Equation as an Approximation for Nonlinear Relationship

assumptions that the conditional distributions of Y are (1) normal, with (2) identical standard deviation σ for each X-value. These assumptions are also never *exactly* satisfied in practice. However, the closer reality falls to this ideal model, the more appropriate are the confidence interval and test procedures introduced in that section. For large samples, the normality assumption is not crucial, since an extended Central Limit Theorem implies that sample slopes and correlations have approximately normal sampling distributions.

Chapter 14 discusses ways of checking the assumptions of regression modeling 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 of X. The relationship might not be linear outside that range. We might get poor or even absurd predictions by extrapolating beyond the observed range.

To illustrate, the prediction equation $\hat{Y} = -.86 + .58X$ in Section 9.2 relating X = poverty rate to Y = murder rate was based on sample poverty rates between 8.0 and 26.4. It is not valid to extrapolate much below or above this range. For instance, the predicted murder rate for a poverty rate of $X = 0\%$ is $\hat{Y} = -.86$. 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 fit of the model. A single observation can have a large effect on the fit if its X value is unusually large or unusually small and if it falls 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, taken from Table 9.13 in Problem 9.17, 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 $\hat{Y} = 29.8 - .024X$ to $\hat{Y} = 31.2 - .195X$. Adding only a single point to the data set causes the prediction line to tilt dramatically downward.

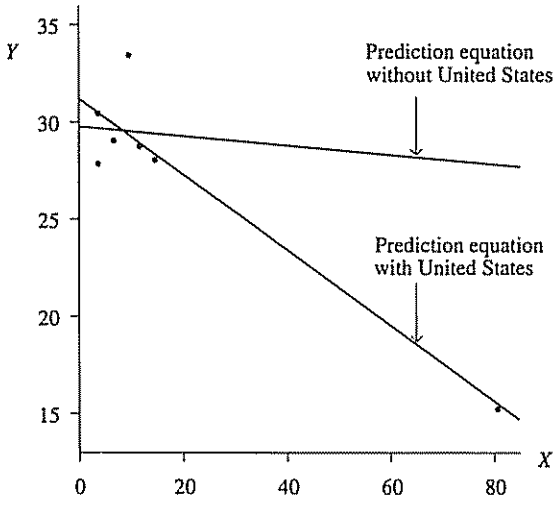

Figure 9.17 Prediction Equations for Y = Birth Rate and X = Television Ownership, with and without Observation for United States

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.

When a scatter diagram reveals a severe outlier, the reasons for it should be investigated. 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. It is often worthwhile to refit the model without one or two extreme outliers to see if those observations have a large effect on the parameter estimates, as we did in Example 9.4 with the D.C. observation for the murder rate data.

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 $-.935$ when the outlier is included and $-.051$ 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 sample correlation depends on the range of X-values sampled. When a sample has a much narrower range of variation in X than the population, for example, the sample correlation tends to underestimate drastically (in absolute value) the population correlation.

Figure 9.18 shows a scatter diagram of 500 points that is very regular and has a correlation of $r = .705$. 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 $r = .330$, considerably lower. For another example, consider the relation between housing price and size of house, portrayed in Figure 9.15. The correlation equals .899. If we sampled only those sales in which house size is between 1.3 and 2.0 thousand feet, which include 48 of the 93 observations, the correlation decreases to .564.

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.

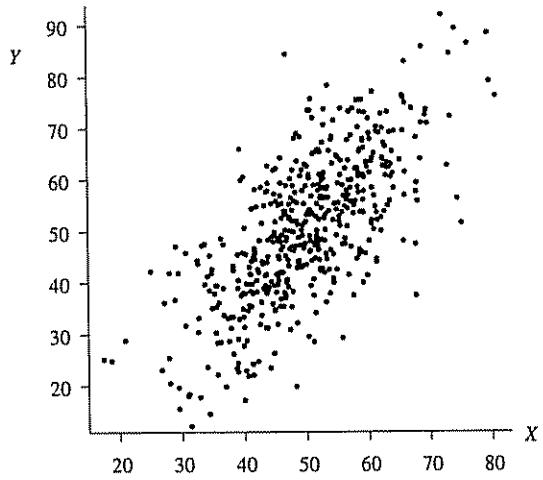

Figure 9.18 The Correlation is Affected by the Range of X Values; Correlation Decreases from .705 to .330 Using Only Points with X Between 43 and 57

Example 9.12 Does the SAT Predict College GPA?

In studying the association between score on a college entrance exam (such as the SAT) and later performance in college (such as GPA at end of second year), 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 not be strong, 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, and 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 .97 for the full data and .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*

We now show an alternative way of writing the linear regression equation. Recall that at a fixed value of X, the values of Y fluctuate around their mean, $E(Y) = \alpha + \beta X$. Any specific observation on Y may fall above that mean (i.e., above the regression line) or below that mean (below the regression line). The standard deviation component σ of the model 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. One does this by including a term for the deviation of the observation from the mean. This approach models an observation on Y as

$$Y = \alpha + \beta X + \varepsilon$$

The term denoted by ε (the Greek letter, epsilon) represents the deviation of the observation 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.

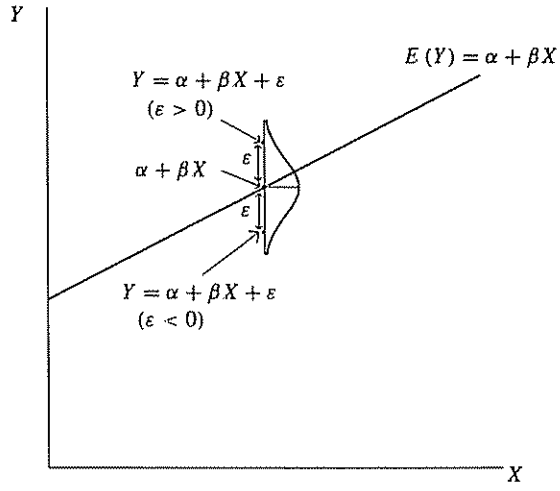

Figure 9.19 Positive and Negative ε-Values Correspond to Observations Above and Below the Conditional Mean

The ε term in the equation $Y = \alpha + \beta X + \varepsilon$ occurs because observations with the same X-values do not all have the same Y-values; that is, the variables are not perfectly correlated. The model $Y = \alpha + \beta X + \varepsilon$ is a *probabilistic* one. 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 for the prediction of the individual observation on Y.

For a particular sample of n observations, 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 simply 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.

One can write the regression model either as $E(Y) = \alpha + \beta X$ or as $Y = \alpha + \beta X + \varepsilon$. Although we use the first equation, we introduced this alternative one since some books and articles express the regression model in this way.

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. However, by definition, models merely approximate reality.

If the model seems too simple to be adequate, the scatter diagram or other diagnostics may suggest improvement by using more general models introduced in the remainder of 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 be easily interpreted. Chapters 11–15 present ways of extending 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. The explanatory variable is a binary variable defining the two groups; we compare means when the response variable is quantitative and proportions when the response variable is qualitative. When the variables are statistically independent, the means or proportions are identical for the two groups.

Chapter 8 dealt with *association between two qualitative 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 regression model is appropriate when a straight line approximates the relationship between X and Y.
- A ***scatter diagram*** plots the data and checks whether the relationship is approximately linear. If it is, the ***method of least squares*** provides estimates of the Y-intercept α and the slope β for a linear prediction equation $\hat{Y} = a + bX$.
- The ***Pearson correlation r*** and its square, the ***coefficient of determination*** r^2, describe the *strength* of the linear association between two quantitative variables. The Pearson correlation is a standardized slope, having the same sign but falling between -1 and $+1$. Its square has a proportional reduction in error (PRE) interpretation; it describes the reduction in variability about the prediction equation compared to the variability about the sample mean of Y.
- For inference about the *existence* of a relationship between the variables, a t test using the slope or correlation tests the ***null hypothesis of independence***, namely, that the population slope and correlation equal 0. More informatively, one can construct confidence intervals for the unknown parameters.

Table 9.8 summarizes the methods studied in the past three chapters.

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	$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$	$z = \frac{\hat{\gamma}}{\sigma_{\hat{\gamma}}}$	$t = \frac{b}{\sigma_b} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$, $df = n - 2$
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		$r^2 = \frac{E_1 - E_2}{E_1} = \frac{\text{TSS}-\text{SSE}}{\text{TSS}}$

Chapter 11 introduces the ***multiple regression*** model, a generalization that permits *several* explanatory variables in the model. Chapter 12 shows how to include qualitative predictors in a regression model, and Chapter 13 includes both qualitative and quantitative predictors. Chapter 14 introduces models for more complex relationships,

such as nonlinear ones. Finally, Chapter 15 presents regression models for qualitative 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.

PROBLEMS

Practicing the Basics

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.
2. Sketch plots of the following lines, for values of X between 0 and 10:
   a) $Y = 7 + .5X$
   b) $Y = 7 + X$
   c) $Y = 7 - X$
   d) $Y = 7 - .5X$
   e) $Y = 7$
   f) $Y = X$
3. For the data on the 50 states in Table 9.1 on Y = violent crime rate and X = poverty rate, the prediction equation is $\hat{Y} = 209.9 + 25.5X$.
   a) Sketch a plot of the prediction equation for X between 0 and 100.
   b) Interpret the Y-intercept and the slope.
   c) Find the predicted violent crime rate for Massachusetts, which has $X = 10.7$ and $Y = 805$.
   d) Find the residual for the Massachusetts prediction. Interpret.
   e) Two states differ by 10.0 in their poverty rates. Find the difference in their predicted violent crime rates.
   f) The state poverty rates range from 8.0 (for Hawaii) to 24.7 (for Mississippi). Over this range, find the range of predicted values for violent crime rate.
   g) What is the sign of the Pearson correlation between these variables? Why?
4. A college admissions officer claims that the prediction equation $\hat{Y} = .5 + 7.0X$ approximates the relationship between Y = college GPA and X = high school GPA (both measured on a four-point scale) for students at that college.
   a) Is this equation realistic? Why or why not?
   b) Suppose that the prediction equation is actually $\hat{Y} = .5 + .7X$. Interpret the slope.
   c) Using the prediction equation in (b), find the predicted GPA for a student having a high school GPA of (i) 3.0, (ii) 4.0.
   d) Suppose the prediction equation is $\hat{Y} = X$. Identify the Y-intercept and slope, and interpret their values.
5. A recent study of mail survey response rate patterns of the elderly found a prediction equation relating X = age and Y = percentage of subjects responding of $\hat{Y} = 90.2 - .6X$, for ages between about 60 and 90 (D. Kaldenberg et al., *Public Opinion Quarterly*, Vol 58, 1994, p. 68).

a) Interpret the slope.
b) Find the predicted response rate for a (i) 60 year old, (ii) 90 year old.
c) Find the difference in predicted response rates for two age groups that are ten years apart.

6. Refer to Problem 9.24. For those counties, Table 9.9 shows part of the printout for the regression analysis relating Y = median income (thousand of dollars) to X = percent of residents with at least a high school education.
a) Report the prediction equation, and interpret the slope.
b) Is the Y-intercept meaningful? Explain.
c) Find the predicted median income for a county with 70% of the residents having at least a high school education.
d) County A has 10% more of its residents than county B with at least a high school education. Find their difference in predicted median incomes.
e) Find the Pearson correlation. Interpret using (i) the sign, (ii) the magnitude, (iii) the standardized slope.
f) Find the coefficient of determination. Explain its PRE interpretation.

TABLE 9.9

Variable	Mean	Std Dev
INCOME	24.51	4.69
EDUCATIO	69.49	8.86

Variable	Parameter Estimate
INTERCEP	-4.63
EDUCATIO	0.42

7. A high school student analyzes whether a relationship exists between X = number of books read for pleasure in the previous year and Y = daily average number of hours spent watching television. For her three best friends, the observations are as shown in Table 9.10.
a) Construct a scatter diagram. From inspection of the diagram, state the prediction equation, and interpret. (Note: You can do this without using the least squares formulas.)
b) Report the sample correlation between X and Y, and interpret.

TABLE 9.10

x	y
0	5
5	3
10	1

8. For the WWW data set described in Problem 1.7, the sample correlation between Y = political ideology (scored 1 to 7) and X = number of times a week reading a newspaper is $r = -.066$.
a) Interpret the sign of the correlation.
b) Interpret the square of the correlation. Would you conclude that the sample association is strong, or weak?
c) When Y is predicted using X = religiosity (how often attend religious services, scored 0, 1, 2, 3), the sample correlation is $r = .580$. Which of these two explanatory variables seems to have a stronger linear relationship with Y? Explain.

9. The prediction equation for a sample of 100 people relating X = years of education and Y = annual income (in dollars) is $\hat{Y} = -6000 + 3000X$, and the Pearson correlation equals .50.

a) Suppose instead that Y refers to annual income, in thousands of dollars. State the prediction equation and the correlation.

b) Suppose that Y is treated as the explanatory variable and X is treated as the response variable. Will the correlation coefficient or the slope change in value? Explain.

10. For the house sales data in Table 9.4, Table 9.11 shows a computer printout for the regression analysis relating selling price (thousands of dollars) to number of bedrooms.

TABLE 9.11

Variable	N	Mean	Std Dev
PRICE	93	99.5333	44.1841
BEDROOMS	93	3.1828	0.6070

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	62577.52	62577.52	48.66	0.0001
Error	91	117028.35	1286.03		
C Total	92	179605.87			

Root MSE 35.86120 R-square 0.3484

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-37.229	19.955	-1.866	0.0653
BEDROOMS	1	42.969	6.160	6.976	0.0001

a) Report the prediction equation, and interpret the slope. Is the relationship positive, or negative?

b) Find the predicted selling price for a home with (i) two, (ii) three, (iii) four bedrooms.

c) The first observation in the data set has three bedrooms and a selling price of 48.5 thousand dollars. Find the residual, and interpret.

d) Using the sample slope and the standard deviations, find the Pearson correlation. Interpret its value.

e) Report the coefficient of determination, and interpret its value.

f) Report the standard error of the sample slope. Interpret.

g) Find the test statistic and P-value for testing H_0: $\beta = 0$ against H_a: $\beta \neq 0$, and interpret.

h) Construct a 95% confidence interval for β, and interpret.

i) Use the result of the previous part to form a 95% confidence interval for the difference in the mean housing prices for homes with $X = 4$ bedrooms and with $X = 2$ bedrooms. Interpret.

j) Interpret the value labeled "Root MSE."

11. Refer to Table 9.1. Table 9.12 shows an SPSS printout for the relationship for all 51 observations between Y = murder rate and X = percentage white.

a) Report the prediction equation. Interpret the Y-intercept and slope.

b) Report the coefficient of determination, and interpret.

TABLE 9.12

R Square		.4988			
Standard Error		7.6646			
Variable	B	SE B	Beta	T	Sig T
WHITE	-0.571	0.0817	-0.706	-6.983	0.0001
Constant	56.766	6.9623		8.153	0.0001

c) Find the correlation, and interpret.
d) Report and interpret the estimated conditional standard deviation of murder rate.
e) When the D.C. observation is deleted, the estimated slope changes to $-.216$, and r^2 changes to .359. Explain how a single observation can have such a large effect.

12. Refer to Table 9.1. For all 51 observations, use software to analyze the relationship between murder rate and violent crime rate, treating murder rate as the response variable.
a) Construct a scatter diagram. Does there seem to be a positive, or a negative, relationship?
b) Find the prediction equation, and interpret the coefficients.
c) Find the predicted murder rate and the residual for D.C. Interpret.
d) Find the predicted murder rate at the mean of 612.8 for violent crime rate.
e) Using the slope and the standard deviations 441.1 for violent crime rate and 10.7 for murder rate, find the correlation. Interpret.
f) Report TSS and SSE from your printout, and use them to find the coefficient of determination and the correlation.
g) Now, treating violent crime rate as the response variable, find the prediction equation. Interpret the coefficients.
h) Find the predicted violent crime rate at the mean of 8.73 for murder rate. (Note from this and from (d) that the predicted value at the mean of the explanatory variable is simply the mean of the response variable.)
i) Using the standard deviations and the slope of this second prediction equation, find the correlation. Compare to (e). How would you characterize this association, strong or weak?
j) Based on box plots for the individual variables or the scatter diagram, would you regard D.C. as an outlier? Refit the model in (b) without it, and note the effect on the slope and correlation.

13. Refer to Problem 9.24. For those data, use software to analyze Y = crime rate and X = percentage living in an urban environment.
a) Construct a stem and leaf plot and a box plot for Y. Interpret.
b) Show that $\hat{Y} = 24.5 + .56X$. Interpret the Y-intercept and slope.
c) Find the predicted crime rate and the residual for Alachua County. Interpret.
d) Using the slope, find the difference in predicted crime rates between counties that are 100% urban and counties that are 0% urban. Interpret.
e) Report and interpret the Pearson correlation. Show the connection between it and the slope and the standard deviations of 28.3 for crime rate and 34.0 for percentage urban.
f) Find TSS and SSE on your printout, and use them to verify the coefficient of determination. Interpret.
g) Does it make sense to conduct statistical inference, such as a test of independence, for these data? Explain.

14. Using software, plot the relationship between percentage single-parent families and percentage white, for the data in Table 9.1.

a) Based on your plot, identify the two observations that seem quite different from the others.

b) Find the prediction equation and the correlation (i) for the entire data set, (ii) deleting the first of the two outlying observations, (iii) deleting the second of the two outlying observations, (iv) deleting both outlying observations. Discuss the influence of these points.

15. Refer to the housing data in Table 9.4. Use software to fit the model with number of bathrooms as the predictor for price.

a) Construct a scatter diagram. Note the effect of using such a highly discrete predictor, having only three values.

b) Find the prediction equation. Interpret the slope.

c) Find the predicted selling price for homes with one, two, and three bathrooms.

d) Find the Pearson correlation and interpret its value.

e) Find the coefficient of determination and interpret its value.

f) Test the null hypothesis that mean selling price is independent of number of bathrooms, and report the P-value. Why is this inference not especially informative for these variables?

g) Find a 95% confidence interval for the difference between the mean selling price for homes with two bathrooms and homes with one bathroom. Interpret.

h) Find a 95% confidence interval for the difference between the mean selling price for homes with three bathrooms and homes with one bathroom. Interpret.

16. A study was conducted using 49 Catholic female undergraduates at Texas A & M University. The variables measured refer to the parents of these students. The response variable is the number of children that the parents have. One of the explanatory variables is the mother's educational level, measured as the number of years of formal education. For these data, $\bar{X} = 9.88, s_X = 3.77, \bar{Y} = 3.35, s_Y = 2.19$, the prediction equation is $\hat{Y} = 5.40 - .207X$, the standard error of the slope estimate is .079, and SSE = 201.95.

a) Interpret the Y-intercept and slope.

b) Find the predicted numbers of children for women with (i) 8, (ii) 12, (iii) 16 years of education.

c) Find the Pearson correlation and interpret its value.

d) Test the null hypothesis that mean number of children is independent of mother's educational level, and report and interpret the P-value.

e) Find a 95% confidence interval for the slope of the regression equation. Interpret.

f) Construct and interpret a 95% confidence interval for the difference between the mean number of children for two sets of mothers who are eight years apart in educational level.

g) Sketch a potential scatter diagram for these variables such that the analyses you conducted above would be inappropriate.

17. Table 9.13 lists recent values for several nations on the crude birth rate (number of births per 1000 population size), women's economic activity (female labor force as percentage of male), percentage women using contraception, female life expectancy, female adult literacy rate, a human development index (HDI, which has components referring to life expectancy at birth, educational attainment, and income per capita), gross national product (GNP, per capita, in thousands of dollars), daily newspaper circulation per 100 people, and number of televisions per 100 people. This exercise uses birth rate as the response

TABLE 9.13

Nation	Birth Rate	Econ. Activ.	Contra-ception	Liter. Rate	Life Expec.	HDI	GNP	News Circul.	TV
Algeria	29.0	11	47	73	68	44	1.6	5	7
Argentina	19.5	38	–	88	76	96	4.0	12	22
Australia	14.1	61	76	93	81	99	16.6	25	48
Brazil	21.2	38	66	80	69	81	2.6	5	21
Canada	13.7	63	–	95	81	99	20.8	23	64
China	17.8	81	83	59	70	70	1.3	5	3
Cuba	14.5	50	70	77	77	94	1.6	17	16
Denmark	12.4	77	78	92	78	99	24.2	35	54
Egypt	28.7	12	46	61	65	36	.5	6	12
France	13.0	64	81	93	81	99	24.1	21	41
Germany	11.0	–	75	92	79	99	19.8	59	56
India	27.8	34	43	44	60	35	.3	3	4
Iraq	43.6	29	14	62	67	41	.7	4	7
Israel	20.4	49	–	91	78	95	13.6	26	27
Japan	10.7	64	64	94	82	99	27.3	59	61
Malaysia	28.0	55	48	75	73	82	2.5	14	15
Mexico	26.6	37	53	84	74	86	3.1	13	15
Nigeria	43.3	51	6	41	52	42	2	–	–
Pakistan	41.8	16	12	48	63	63	.4	2	2
Philippines	30.4	44	40	68	68	94	.7	5	4
Russia	12.6	70	–	85	74	99	8.6	–	–
South Africa	33.4	54	50	–	66	70	2.6	4	10
Spain	11.2	31	–	98	80	93	13.4	8	40
United Kingdom	13.2	60	81	92	79	99	17.4	39	43
United States	15.2	65	74	94	79	99	22.6	25	81
Vietnam	26.3	82	53	89	67	54	–	–	–

Sources: *Statistical Abstract of the United States, 1995* (Washington: D.C.: U.S. Government Printing Office), and *Human Development Report, 1995* (Oxford: Oxford University Press, 1995).

variable and women's economic activity as the explanatory variable. Table 9.14 shows part of a SPSS printout for a regression analysis.

a) Report the prediction equation, and interpret the Y-intercept and slope.

b) Report r and r^2, and interpret their values.

c) Find the predicted value and residual for Nigeria, and interpret.

18. Refer to the previous exercise. Now use GNP as the explanatory variable for predicting birth rate.

a) Construct a stem and leaf plot or box plot for birth rate, and describe its distribution.

b) Construct a scatter diagram, and indicate whether a linear model seems appropriate.

c) Fit the model, and interpret the parameter estimates.

d) Can you compare the slopes of the prediction equations with the two predictors to determine which has the stronger effect? Explain.

e) Which variable, GNP or women's economic activity, seems to have the stronger association with birth rate?

TABLE 9.14

```
R Square                .2973
Standard Error         8.821

Variable          B        SE B     Beta       T        Sig T
WOMEN-EC       -0.277      0.089   -0.545    -3.119     0.0048
(Constant)     36.324      4.700              7.729     0.0001
```

19. Refer to the previous two exercises. Using software, obtain the correlation matrix for these data. Which pairs of variables are highly correlated? Describe the nature of those correlations, and explain how your software handled the missing values. (For a particular analysis, most software deletes observations for which data are missing on at least one variable used in the analysis. Better strategies exist; see, for instance, R. Little and D. Rubin, *Sociological Methods and Research*, Vol. 18, 1989, pp. 292–326.)

20. For a random sample of U.S. counties, data are obtained on X = percentage of the population aged over 50 and Y = per student expenditure on education. Table 9.15 is part of the computer printout for the analysis.
a) What was the sample size for this study?
b) Fill in the blanks in Table 9.15.

TABLE 9.15

Source	DF	Sum of Squares	Mean Square	Root MSE	R-square
Model	1	-------	-------	-----	-----
Error	--	400000	-------		
Total	49	500000			

Variable	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1300.0	6.334	205.2	.0000
AGE	-5.0	1.443	-----	-----

21. For Table 9.1, use software to analyze the data on violent crime rate and percent single parent families.
a) Construct a scatter diagram. What does it show?
b) One point is quite far removed from the others, having a much higher value on both variables than the rest of the sample, but it fits in well with the linear trend exhibited by the rest of the points. Show that the correlation changes from .839 to .649 when you delete this observation. Why does it drop so dramatically?

Concepts and Applications

22. Refer to the WWW data set (Problem 1.7). Using software, conduct regression analyses relating (i) Y = political ideology and X = religiosity, (ii) Y = high school GPA and X = hours of TV watching. Prepare a report,
a) Using graphical ways of portraying the individual variables and their relationship.

b) Interpreting descriptive statistics for summarizing the individual variables and their relationship.

c) Summarizing and interpreting results of inferential analyses.

d) Checking effects of possibly influential outliers.

23. Refer to the data file you created in Problem 1.7. For variables chosen by your instructor, conduct a regression and correlation analysis. Report both descriptive and inferential statistical analyses, interpreting and summarizing your findings.

24. Table 9.16 shows recent data from all 67 Florida counties on crime rate (number of crimes per 1000 residents), median income (in thousands of dollars), percentage of residents with at least a high school education (of those aged at least 25), and the percentage of the county's residents living in an urban environment. Using crime rate as the response vari-

TABLE 9.16

County	Crime Rate	Median Income	High School	Percent Urban	County	Crime Rate	Median Income	High School	Percent Urban
ALACHUA	104	22.1	82.7	73.2	LAFAYETTE	0	20.7	58.2	0.0
BAKER	20	25.8	64.1	21.5	LAKE	42	23.4	70.6	43.2
BAY	64	24.7	74.7	85.0	LEE	59	28.4	76.9	86.1
BRADFORD	50	24.6	65.0	23.2	LEON	107	27.3	84.9	82.5
BREVARD	64	30.5	82.3	91.9	LEVY	45	18.8	62.8	0.0
BROWARD	94	30.6	76.8	98.9	LIBERTY	8	22.3	56.7	0.0
CALHOUN	8	18.6	55.9	0.0	MADISON	26	18.2	56.5	20.3
CHARLOTTE	35	25.7	75.7	80.2	MANATEE	79	26.0	75.6	88.7
CITRUS	27	21.3	68.6	31.0	MARION	64	22.5	69.6	39.6
CLAY	41	34.9	81.2	65.8	MARTIN	53	31.8	79.7	83.2
COLLIER	55	34.0	79.0	77.6	MONROE	89	29.4	79.7	73.2
COLUMBIA	69	22.0	69.0	31.1	NASSAU	42	30.2	71.2	44.9
DADE	128	26.9	65.0	98.8	OKALOOSA	37	27.9	83.8	84.0
DESOTO	69	21.0	54.5	44.6	OKEECH	51	21.4	59.1	30.1
DIXIE	49	15.4	57.7	0.0	ORANGE	93	30.3	78.8	93.1
DUVAL	97	28.5	76.9	98.8	OSCEOLA	78	27.3	73.7	66.4
ESCAMBIA	70	25.2	76.2	85.9	PALM B.	90	32.5	78.8	94.7
FLAGLER	34	28.6	78.7	63.1	PASCO	42	21.5	66.9	67.4
FRANKLIN	37	17.2	59.5	30.2	PINELLAS	70	26.3	78.1	99.6
GADSDEN	52	20.0	59.9	28.8	POLK	84	25.2	68.0	70.3
GILCHRIST	15	20.6	63.0	0.0	PUTNAM	83	20.2	64.3	15.7
GLADES	62	20.7	57.4	0.0	SANTA R.	43	27.6	79.9	57.2
GULF	19	21.9	66.4	35.2	SARASOTA	58	29.9	71.7	92.1
HAMILTON	6	18.7	58.4	0.0	SEMINOLE	56	35.6	78.5	44.4
HARDEE	57	22.1	54.8	16.7	ST JOHNS	54	29.9	81.3	93.2
HENDRY	47	24.9	56.6	44.7	ST LUCIE	58	27.7	84.6	92.8
HERNANDO	44	22.7	70.5	61.3	SUMTER	37	19.6	64.3	19.3
HIGHLANDS	56	21.1	68.2	24.8	SUWANEE	37	19.8	63.8	23.6
HILLSBOR	110	28.5	75.6	89.2	TAYLOR	76	21.4	62.1	41.8
HOLMES	5	17.2	57.1	16.8	UNION	6	22.8	67.7	0.0
INDIAN R.	58	29.0	76.5	83.0	VOLUSIA	62	24.8	75.4	83.9
JACKSON	32	19.5	61.6	21.7	WAKULLA	29	25.0	71.6	0.0
JEFFERSON	36	21.8	64.1	22.3	WALTON	18	21.9	66.5	20.9
					WASHINGTON	21	18.3	60.9	22.9

Source: Dr. Larry Winner, University of Florida

able and percent urban as the predictor, analyze these data. In your report, provide interpretations of all the analyses.

25. Refer to the previous exercise. Using income as the response variable and percentage of high school graduates as the explanatory variable, analyze these data. Prepare a report, and explain carefully the interpretations of all your analyses.

26. Refer to Table 9.1. Analyze the relationship between violent crime rate and percentage having at least a high school education. Write a report showing your analyses, providing interpretations, and summarizing your findings.

27. Refer to Table 9.1. Analyze the relationship between violent crime rate and percentage of single-parent families. Write a report showing your analyses, providing interpretations, and summarizing your findings.

28. Repeat the previous exercise, using murder rate as the response variable.

29. Refer to Table 9.13. Analyze the relationship between newspaper circulation and gross national product. Tell why you conducted each part of the analysis and explain how to interpret the results.

30. Recently the General Social Survey has asked subjects to rate various groups using the "feeling thermometer." Ratings between 50 and 100 mean you feel favorable and warm toward the group, whereas ratings between 0 and 50 mean that you don't feel favorable. It also asks subjects to rate themselves on political ideology, using scores 1 = extremely liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = extremely conservative, and to describe their religious attendance, using the categories (never, less than once a year, once or twice a year, several times a year, about once a month, 2-3 times a month, nearly every week, every week, several times a week). Table 9.17 shows data for ten of the subjects in a recent survey, where the feelings thermometer refers to feelings about liberals and using religion scores that are the category numbers.

a) Analyze Table 9.17. Tell why you conducted each analysis, and explain how to interpret the results.

d) Suppose the feelings response for the first subject had incorrectly been recorded as 90 instead of 10. How would this have affected results of your analyses?

TABLE 9.17

Feelings	Ideology	Religion	Feelings	Ideology	Religion
10	7	9	60	4	8
70	3	5	40	5	6
50	4	7	80	3	7
30	6	4	50	4	6
90	4	3	40	6	8

31. In an article in *USA Today* (December 28, 1984), sociologists N. Glenn and B. A. Shelton are quoted as showing a strong link between residential mobility and divorce rates. In Table 9.18, divorce rate is the annual number of divorces and annulments per 1000 population, and mobility rate is the percentage of people living in a different house from five years ago. Analyze these data.

32. Describe a situation in which it is inappropriate to use the Pearson correlation to measure the association between two quantitative variables.

TABLE 9.18

Region	Mobility Rate	Divorce Rate
New England	41	4.0
Middle Atlantic	37	3.4
East North Central	44	5.1
West North Central	46	4.6
South Atlantic	47	5.6
East South Central	44	6.0
West South Central	50	6.5
Mountain	57	7.6
Pacific	56	5.9

33. Annual income, in dollars, is the response variable in a regression analysis. For a British version of the report on the analysis, all responses are converted to British pounds sterling (1 pound equals about 1.5 dollars, as of 1997).
a) How, if at all, does the slope of the prediction equation change?
b) How, if at all, does the correlation change?

34. The variables Y = annual income (thousands of dollars), X_1 = number of years of education, and X_2 = number of years experience in job are measured for all the employees having city-funded jobs, in Knoxville, Tennessee. The following prediction equations and correlations apply.
i. $\hat{Y} = 10 + 1.0X_1$, $r = .30$
ii. $\hat{Y} = 14 + .4X_2$, $r = .60$.
The correlation is $-.40$ between X_1 and X_2. Which of the following statements are true?
a) The strongest sample association is between Y and X_2.
b) The weakest sample association is between X_1 and X_2.
c) The prediction equation using X_2 to predict X_1 has negative slope.
d) A standard deviation increase in education corresponds to a predicted increase of .3 standard deviations in income.
e) There is a 30% reduction in error in using education, instead of $\bar{Y}$, to predict income.
f) Each additional year on the job corresponds to a $400 increase in predicted income.
g) When X_1 is the predictor of Y, the sum of squared residuals (SSE) is larger than when X_2 is the predictor of Y.
h) The predicted mean income for employees having 20 years of experience is $4000 higher than the predicted mean income for employees having 10 years of experience.
i) If $\hat{\sigma} = 8$ for the model using X_1 to predict Y, then it is not unusual to observe an income of $70,000 for an employee who has 10 years of education.
j) It is possible that $s_Y = 12.0$ and $s_{X_1} = 3.6$.
k) It is possible that $\bar{Y} = 20$ and $\bar{X}_1 = 13$.

Select the best response(s) in Problems 7.35–7.37.

35. One can interpret $r = .3$ as follows:
a) A 30% reduction in error occurs in using X to predict Y.
b) A 9% reduction in error occurs in using X to predict Y compared to using $\bar{Y}$ to predict Y.

c) 9% of the time $\hat{Y} = Y$.
d) Y changes .3 unit for every one-unit increase in X.
e) When X predicts Y, the average residual is .3.
f) X changes .3 standard deviations when Y changes one standard deviation.

36. The correlation is inappropriate as a measure of association between two quantitative variables:
a) When different people measure the variables using different units.
b) When the relationship is highly nonlinear.
c) When the data points fall exactly on a straight line.
d) When the slope of the prediction equation is 0 using nearly all the data, but a couple of outliers are extremely high on Y at the high end of the X scale.
e) When Y tends to decrease as X increases.
f) When we have data for the entire population rather than a sample.
g) When the sample has a much narrower range of X-values than does the population.

37. The slope of the least squares prediction equation and the Pearson correlation coefficient are similar in the sense that
a) They do not depend on the units of measurement.
b) They both must fall between -1 and $+1$.
c) They both have the same sign.
d) They both equal 1 when there is the strongest association
e) Their squares both have PRE interpretations.
f) They have the same t statistic value for testing H_0: Independence.
g) They both can be strongly affected by severe outliers.

38. Describe the assumptions (a) in using the regression equation $E(Y) = \alpha + \beta X$ to represent the relationship between two variables and (b) in making inferences about that equation using the least squares prediction equation. Which assumptions are most critical?

39. Refer to the previous exercise. In view of these assumptions, indicate why such a model would or would not be good in the following situations.
a) X = time, Y = percentage unemployed workers in the United States. (*Hint*: Does this continually tend to increase or decrease?)
b) X = income, Y = charitable contributions within the previous year. (*Hint*: Would poor people show as much variation as wealthy people?)
c) X = age, Y = annual medical expenses. (*Hint*: Suppose expenses tend to be relatively high for the newborn and for the elderly.)
d) X = per capita income, Y = life expectancy, for nations. (*Hint*: The increasing trend eventually levels off.)

40. For a class of 100 students, the teacher takes the 10 students who perform poorest on the midterm exam and enrolls them in a special tutoring program. The overall class mean is 70 both on the midterm and final, but the mean for the specially tutored students increases from 50 to 60. Can we conclude that the tutoring program was successful? Explain.

41. Refer to Problem 9.24. For these counties, the correlation between high school education rate and income equals .79. Suppose we also have data at the individual level as well as aggregated for a county. Sketch a scatter diagram to show that at the individual level, the correlation could be much weaker. (*Hint*: Show that lots of variability could exist for individuals, yet the summary values for counties could fall close to a straight line.) Because of this result, it is misleading to extend results from the aggregate level to indi-

viduals. Making predictions about individuals based on the behavior of aggregate groups is known as the ***ecological fallacy*** (See W. S. Robinson, *American Sociological Review*, Vol. 15, 1950, p. 351).

42. Explain why the correlation between X = number of years of education and Y = annual income is likely to be smaller if we use a random sample of adults who have a college degree than if we use a random sample of all adults.

43 Explain carefully the interpretations of the standard deviations (a) s_Y, (b) s_X, (c) $\hat{\sigma}$, (d) $\hat{\sigma}_b$.

44. *A report summarizing the results of a study on the relationship between scores for students on a verbal aptitude test X and a mathematics aptitude test Y states that $\bar{X} = 480$, $\bar{Y} = 500$, $s_X = 80$, $s_Y = 120$, and $r = .60$.

a) Using the formulas for the correlation and for the least squares estimates, find the prediction equation.

b) Find the prediction equation for predicting verbal test result using math test result.

45 *Observations on both X and Y are standardized, having estimated means of 0 and standard deviations of 1 (see Section 4.2). Show that the prediction equation has the form $\hat{Y} = rX$, where r is the sample correlation between X and Y; that is, for the standardized variables, the Y-intercept equals 0 and the slope is the same as the correlation.

46. *A confidence interval for a population correlation ρ requires a mathematical transformation of r for which the sampling distribution is approximately normal. This transformation is $T(r) = (1/2)\log[(1 + r)/(1 - r)]$, where *log* denotes the natural (base-e) logarithm. The transformation of the population value ρ is denoted by $T(\rho)$. The variable $T(r)$ is approximately normally distributed about $T(\rho)$ with standard error $\sigma_T = 1/\sqrt{n - 3}$. A confidence interval for $T(\rho)$ is $T(r) \pm z\sigma_T$. Once we get the endpoints of the interval for $T(\rho)$, we substitute each endpoint for T in the inverse transformation $\rho = (e^{2T} - 1)/(e^{2T} + 1)$, where e denotes the exponential function (the inverse of the natural log function). These two values form the endpoints of the confidence interval for ρ.

a) For the correlation of .899 for housing price and size in Table 9.4, show that $T(r) = 1.47$. Show that the standard error of $T(r)$ is .1054.

b) Show that a 95% confidence interval for $T(\rho)$ is (1.26, 1.67).

c) Show that the corresponding confidence interval for ρ is (.85, .93). (Unless $r = 0$, the confidence interval for ρ is not symmetric about the point estimate r, because of the nonsymmetry of the sampling distribution of r.)

d) A confidence interval for the population value ρ^2 of the coefficient of determination follows directly by squaring the limits of the confidence interval for ρ. Find and interpret this confidence interval.

e) If the confidence interval for ρ includes 0, explain why the lower endpoint of the confidence interval for ρ^2 is also 0, and the upper endpoint is the larger of the squared endpoints of the confidence interval for ρ.

47. *Refer to the previous exercise and to Problem 9.16. Find and interpret 95% confidence intervals for the population Pearson correlation and the population coefficient of determination.

48. *Refer to Problem 9.46. Let ρ_1 and ρ_2 denote the population correlation values between two variables for two separate populations. Let r_1 and r_2 denote sample values for inde-

pendent random samples from the populations. To test the null hypothesis H_0: $\rho_1 = \rho_2$, the test statistic is

$$z = \frac{T_2 - T_1}{\sigma_{T_2 - T_1}}$$

where T_1 and T_2 are the transformed values of r_1 and r_2 and

$$\sigma_{T_2 - T_1} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

If H_0 is true, this test statistic has approximately the standard normal distribution. In Table 9.4, the correlation between housing price and size of home is $r_1 = .932$ for the 28 new homes and $r_2 = .915$ for the 65 older homes. Find the P-value for testing H_0: $\rho_1 = \rho_2$ against H_a: $\rho_1 \neq \rho_2$. Interpret.

49. *Show that substituting $X = \bar{X}$ into the prediction equation $\hat{Y} = a + bX$ yields the predicted Y-value of $\hat{Y} = \bar{Y}$. (*Hint*: The least squares formula for the Y-intercept is $a = \bar{Y} - b\bar{X}$.) Show that this means that the least squares prediction equation passes through the point with coordinates $(\bar{X}, \bar{Y})$, the center of gravity of the data.

50. *Alternative formulas for defining the Pearson correlation use the data in formulas similar to the one for b:

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\left[\sum (X - \bar{X})^2\right]\left[\sum (Y - \bar{Y})^2\right]}} = \frac{1}{n-1}\sum\left(\frac{X - \bar{X}}{s_X}\right)\left(\frac{Y - \bar{Y}}{s_Y}\right)$$

Roughly, the correlation is the average cross-product of the z-score for X times the z-score for Y. Using this formula, explain why (a) the correlation has the same value when X predicts Y as when Y predicts X, (b) the correlation does not depend on the units of measurement. (Note: For the population, the correlation is often defined as

$$\frac{\text{Covariance of } X \text{ and } Y}{(\text{Standard deviation of } X)(\text{Standard deviation of } Y)}$$

where the ***covariance*** between X and Y is the average of the cross-products $(X - \mu_X)(Y - \mu_Y)$ about the population means.)

51. *The values of Y are multiplied by a constant c. From their formulas, show that the standard deviation s_Y and the least squares slope b are also then multiplied by c. Thus, show that $r = bs_X/s_Y$ remains the same, so that r does not depend on the units of measurement.

52. *Suppose that the linear regression equation $E(Y) = \alpha + \beta X$ with normality and constant standard deviation σ is truly appropriate for the relationship between Y and X. Then, the interval of numbers

$$\hat{Y} \pm t\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X - \bar{X})^2}}$$

predicts where a new observation on Y will fall at that value of X. This interval is called a ***prediction interval*** for Y. To make an inference about the *mean* of Y (rather than a single value of Y) at that value of X, one can use the ***confidence interval***

$$\hat{Y} \pm t\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X - \bar{X})^2}}$$

The t-value in these intervals is based on $df = n - 2$. Most software has options for calculating these formulas (e.g., the options CLI and CLM in PROC REG in SAS). Refer to the housing data in Table 9.4, at house size $X = 2.0$.

a) Show that $\hat{Y} = 126.0$ and a 95% prediction interval is (87.0, 165.0).

b) Show that a 95% confidence interval for the mean selling price is (121.2, 130.8).

c) Explain intuitively why a prediction interval for a single observation is much wider than a confidence interval for the mean.

d) Results using these formulas are typically overly optimistic, because the model assumptions never hold exactly. Explain how prediction intervals would likely be in error if, in fact, the variability in housing prices tends to increase as house size increases.

53. *Refer to Problem 9.16 and the previous exercise.

a) Construct a 95% confidence interval for the mean number of children for mothers having $X = 16$.

b) Explain why the prediction interval is probably inappropriate. (*Hint*: Is Y approximately normal?)

54. *Refer to Example 9.8 on regression toward the mean. Does this imply that, over time, variation in height is decreasing until, eventually, everyone has the same height? Explain.

55. *To implement least squares, one can find the formulas for the a and b estimates that minimize SSE = $\sum(Y - \hat{Y})^2 = \sum[Y - (a + bX)]^2$ using calculus, by taking the derivative of this function with respect to a, taking the derivative with respect to b, setting the two derivatives equal to 0, and solving the two linear equations simultaneously for a and b. Take the derivative with respect to a and solve for a, showing that $a = \bar{Y} - b\bar{X}$.

Bibliography

Chatterjee, S., Handcock, M., and Simonoff, J. (1995). *A Casebook for a First Course in Statistics and Data Analysis*. New York: Wiley.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*, 2nd ed. New York: Wiley.

Weisberg, S. (1985). *Applied Linear Regression*, 2nd ed. New York: Wiley.

Younger, M. S. (1985). *A First Course in Linear Regression*, 2nd ed. Boston: Duxbury Press.