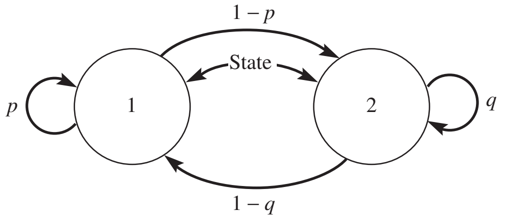
6 Discrete Probabilistic Modeling

Introduction

We have been developing models using proportionality and determining the constants of proportionality. However, what if variation actually takes place as in the demand for gasoline at the gasoline station? In this chapter, we will allow the constants of proportionality to vary in a *random* fashion rather than to be fixed. We begin by revisiting discrete dynamical systems from Chapter 1 and introduce scenarios that have probabilistic parameters.

6.1 Probabilistic Modeling with Discrete Systems

In this section we revisit the systems of difference equations studied in Section 1.4, but now we allow the coefficients of the systems to vary in a probabilistic manner. A special case, called a Markov chain, is a process in which there are the same finite number of states or outcomes that can be occupied at any given time. The states do not overlap and cover all possible outcomes. In a Markov process, the system may move from one state to another, one for each time step, and there is a probability associated with this transition for each possible outcome. The sum of the probabilities for transitioning from the *present state* to the *next state* is equal to 1 for each state at each time step. A Markov process with two states is illustrated in Figure 6.1.



© Cengage Learning

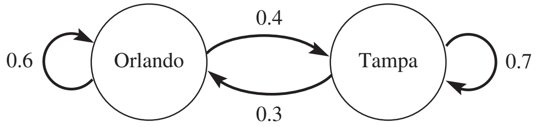
Figure 6.1 A Markov chain with two states; the sum of the probabilities for the transition from a present state is 1 for each state (e.g., $p + (1 - p) = 1$ for state 1).

EXAMPLE 1 *Rental Car Company Revisited*

Consider a rental car company with branches in Orlando and Tampa. Each rents cars to Florida tourists. The company specializes in catering to travel agencies that want to arrange tourist activities in both Orlando and Tampa. Consequently, a traveler will rent a car in one city and drop off the car in either city. Travelers begin their itinerary in either city. Cars can be returned to either location, which can cause an imbalance in available cars to rent. The following historical data on the percentages of cars rented and returned to these locations have been collected for the previous few years.

		Next state	
		Orlando	Tampa
Present state	Orlando	0.6	0.4
	Tampa	0.3	0.7

This array of data is called a **transition matrix** and shows that the probability for returning a car to Orlando that was also rented in Orlando is 0.6, whereas the probability that it will be returned in Tampa is 0.4. Likewise, a car rented in Tampa has a 0.3 likelihood of being returned to Orlando and a 0.7 probability of being returned to Tampa. This represents a Markov process with two states: Orlando and Tampa. Notice that the sum of the probabilities for transitioning from a present state to the next state, which is the sum of the probabilities in each row, equals 1 because all possible outcomes are taken into account. The process is illustrated in Figure 6.2.



© Cengage Learning

Figure 6.2
Two-state Markov chain for the rental car example

Model Formulation Let's define the following variables:

$$p_n = \text{the percentage of cars available to rent in Orlando at the end of period } n$$
$$q_n = \text{the percentage of cars available to rent in Tampa at the end of period } n$$

Using the previous data and discrete modeling ideas from Section 1.4, we construct the following probabilistic model:

$$\begin{aligned} p_{n+1} &= 0.6p_n + 0.3q_n \\ q_{n+1} &= 0.4p_n + 0.7q_n \end{aligned} \qquad (6.1)$$

Model Solution Assuming that all of the cars are originally in Orlando, numerical solutions to System (6.1) give the long-term behavior of the percentages of cars available at each location. The sum of these long-term percentages or probabilities also equals 1. Table 6.1 and Figure 6.3 show the results in table form and graphically:
Notice that

$$\begin{aligned} p_k &\to 3/7 = 0.428571 \\ q_k &\to 4/7 = 0.571429 \end{aligned}$$

Model Interpretation If the two branches begin the year with a total of n cars, then in the long run, approximately 57% of the cars will be in Tampa and 43% will be in Orlando.

Table 6.1 Iterated solution to the rental car example

n	Orlando	Tampa
0	1	0
1	0.6	0.4
2	0.48	0.52
3	0.444	0.556
4	0.4332	0.5668
5	0.42996	0.57004
6	0.428988	0.571012
7	0.428696	0.571304
8	0.428609	0.571391
9	0.428583	0.571417
10	0.428575	0.571425
11	0.428572	0.571428
12	0.428572	0.571428
13	0.428572	0.571428
14	0.428571	0.571429

© Cengage Learning

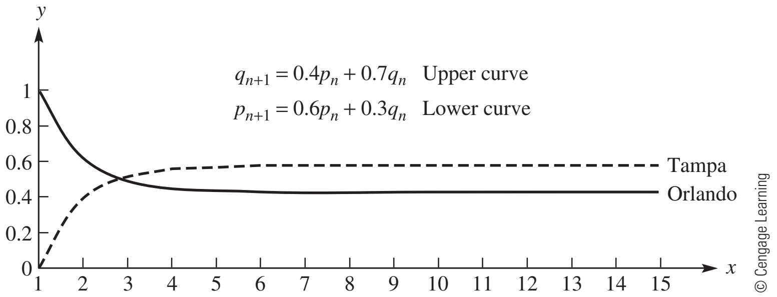


Figure 6.3 Graphical solution to the rental car example

© Cengage Learning

Thus, starting with 100 cars in each location, about 114 cars will be based out of Tampa and 86 will be based out of Orlando in the steady state (and it would take only approximately 5 days to reach this state).

EXAMPLE 2 *Voting Tendencies*

Presidential voting tendencies are of interest every 4 years. During the past decade, the Independent party has emerged as a viable alternative for voters in the presidential race. Let's consider a three-party system with Republicans, Democrats, and Independents.

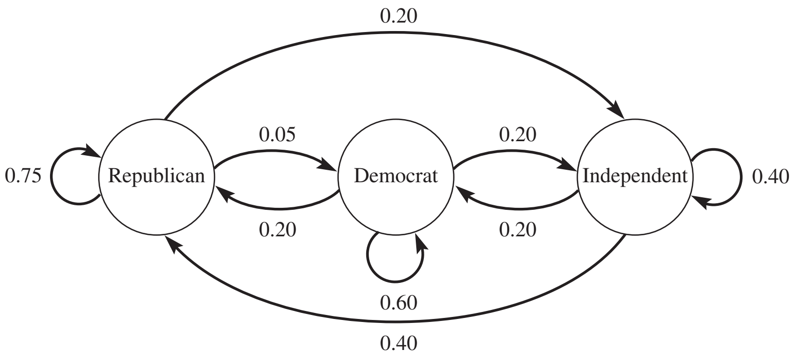
Problem Identification *Can we find the long-term behavior of voters in a presidential election?*

Assumptions During the past decade, the trends have been to vote less strictly along party lines. We provide hypothetical historical data for voting trends in the past 10 years of statewide voting. The data are presented in the following hypothetical transition matrix and shown in Figure 6.4.

		Next state		
		Republicans	Democrats	Independents
Present state	Republicans	0.75	0.05	0.20
	Democrats	0.20	0.60	0.20
	Independents	0.40	0.20	0.40

Figure 6.4

Three-state Markov chain for presidential voting tendencies



© Cengage Learning

Model Formulation Let's define the following variables:

$$
\begin{aligned}
R_n &= \text{the percentage of voters to vote Republican in period } n \\
D_n &= \text{the percentage of voters to vote Democratic in period } n \\
I_n &= \text{the percentage of voters to vote Independent in period } n
\end{aligned}
$$

Using the previous data and the ideas on discrete dynamical systems from Chapter 1, we can formulate the following system of equations giving the percentage of voters who vote Republican, Democratic, or Independent at each time period.

$$
\begin{aligned}
R_{n+1} &= 0.75R_n + 0.20D_n + 0.40I_n \\
D_{n+1} &= 0.05R_n + 0.60D_n + 0.20I_n \\
I_{n+1} &= 0.20R_n + 0.20D_n + 0.40I_n
\end{aligned}
\tag{6.2}
$$

Model Solution Assume that initially 1/3 of the voters are Republican, 1/3 are Democrats, and 1/3 are Independent. We then obtain the numerical results shown in Table 6.2 for the percentage of voters in each group at each period n. The table shows that in the long run (and after approximately 10 time periods), approximately 56% of the voters cast their ballots for the Republican candidate, 19% vote Democrat, and 25% vote Independent. Figure 6.5 shows these results graphically.

Let's summarize the ideas of a Markov chain. A **Markov chain** is a process consisting of a sequence of events with the following properties:

1. An event has a finite number of outcomes, called **states**. The process is always in one of these states.

Table 6.2 Iterated solution to the presidential voting problem

n	Republican	Democrat	Independent
0	0.33333	0.33333	0.33333
1	0.449996	0.283331	0.266664
2	0.500828	0.245831	0.253331
3	0.52612	0.223206	0.250664
4	0.539497	0.210362	0.250131
5	0.546747	0.203218	0.250024
6	0.550714	0.199273	0.250003
7	0.552891	0.1971	0.249999
8	0.554088	0.195904	0.249998
9	0.554746	0.195247	0.249998
10	0.555108	0.194885	0.249998
11	0.555307	0.194686	0.249998
12	0.555416	0.194576	0.249998
13	0.555476	0.194516	0.249998
14	0.55551	0.194483	0.249998

© Cengage Learning

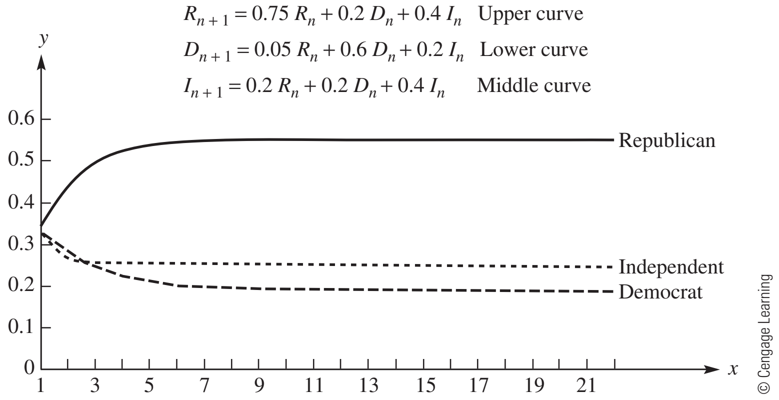


Figure 6.5 Graphical solution to the presidential voting tendencies example

© Cengage Learning

2. At each stage or period of the process, a particular outcome can transition from its present state to any other state or remain in the same state.

3. The probability of going from one state to another in a single stage is represented by a **transition matrix** for which the entries in each row are between 0 and 1; each row sums to 1. These probabilities depend only on the present state and not on past states.

6.1 PROBLEMS

1. Consider a model for the long-term dining behavior of the students at College USA. It is found that 25% of the students who eat at the college's Grease Dining Hall return to eat there again, whereas those who eat at Sweet Dining Hall have a 93% return rate. These are the only two dining halls available on campus, and assume that all students eat at

one of these halls. Formulate a model to solve for the long-term percentage of students eating at each hall.

2. Consider adding a pizza delivery service as an alternative to the dining halls. Table 6.3 gives the transition percentages based on a student survey. Determine the long-term percentages eating at each place.

Table 6.3 Survey of dining at College USA

		Next state		
		Grease Dining Hall	Sweet Dining Hall	Pizza delivery
Present state	Grease Dining Hall	0.25	0.25	0.50
	Sweet Dining Hall	0.10	0.30	0.60
	Pizza delivery	0.05	0.15	0.80

© Cengage Learning

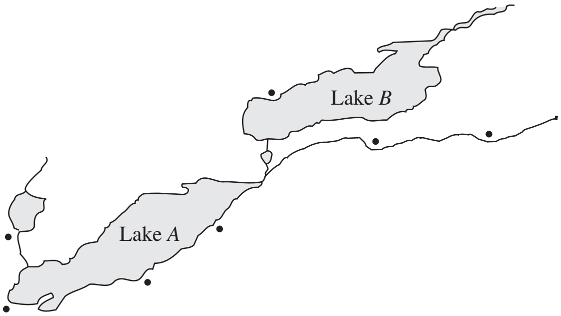
3. In Example 1, assume that all cars were initially in Orlando. Try several different starting values. Is equilibrium achieved in each case? If so, what is the final distribution of cars in each case?

4. In Example 2, it was assumed that initially the voters were equally divided among the three parties. Try several different starting values. Is equilibrium achieved in each case? If so, what is the final distribution of voters in each case?

6.1 PROJECT

1. Consider the pollution in two adjoining lakes in which the lakes are shown in Figure 6.6 and assume the water flows freely between the two lakes but the pollutants flow as in the Markov state diagram, Figure 6.7. Let a_n and b_n be the total amounts of pollution in Lake A and Lake B, respectively, after n years. It has also been determined that we can measure the amount of pollutants given in the lake in which they originated. Formulate and solve the model of the pollution flow as a Markov chain using a system of difference equations.

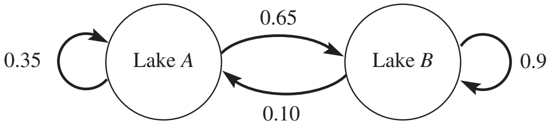
Figure 6.6
Pollution between two lakes



© Cengage Learning

Figure 6.7

Two-state Markov chain for the two lake pollution project.



© Cengage Learning

6.2 Modeling Component and System Reliability

Do your personal computer and your automobile perform well for a reasonably long period of time? If they do, we say that these systems are *reliable*. The **reliability** of a component or system is the probability that it will not fail over a specific time period n. Let's define $f(t)$ to be the failure rate of an item, component, or system over time t, so $f(t)$ is a probability distribution. Let $F(t)$ be the cumulative distribution function corresponding to $f(t)$ as we discussed in Section 5.3. We define the reliability of the item, component, or system by

$$R(t) = 1 - F(t) \tag{6.3}$$

Thus, the reliability at any time t is 1 minus the expected cumulative failure rate at time t.

Human–machine systems, whether electronic or mechanical, consist of components, some of which may be combined to form subsystems. (Consider systems such as your personal computer, your stereo, or your automobile.) We want to build simple models to examine the reliability of complex systems. We now consider design relationships in series, parallel, or combinations of these. Although individual item failure rates can follow a wide variety of distributions, we consider only a few elementary examples.

EXAMPLE 1 *Series Systems*

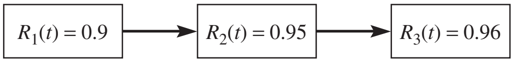
A **series system** is one that performs well as long as *all* of the components are fully functional. Consider a NASA space shuttle's rocket propulsion system as displayed in Figure 6.8. This is an example of a series system because failure for any one of the *independent* booster rockets will result in a failed mission. If the reliabilities of the three components are given by $R_1(t) = 0.90$, $R_2(t) = 0.95$, and $R_3(t) = 0.96$, respectively, then the **system reliability** is defined to be the product

$$R_s(t) = R_1(t)R_2(t)R_3(t) = (0.90)(0.95)(0.96) = 0.8208$$

Note that in a series relationship the reliability of the whole system is less than any single component's reliability, because each component has a reliability that is less than 1.

Figure 6.8

A NASA rocket propulsion system for a space shuttle showing the booster rockets in series (for three stages)



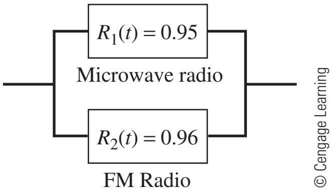
© Cengage Learning

EXAMPLE 2 *Parallel Systems*

A **parallel system** is one that performs as long as a single one of its components remains operational. Consider the communication system of a NASA space shuttle, as displayed in Figure 6.9. Note that there are two separate and *independent* communication systems, either of which can operate to provide satisfactory communications with NASA control. If the independent component reliabilities for these communication systems are $R_1(t) = 0.95$ and $R_2(t) = 0.96$, then we define the **system reliability** to be

$$R_s(t) = R_1(t) + R_2(t) - R_1(t)R_2(t) = 0.95 + 0.96 - (0.95)(0.96) = 0.998$$

Figure 6.9
Two NASA space shuttle communication systems operating in parallel



© Cengage Learning

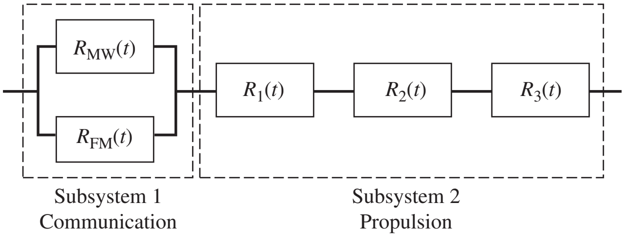
Note that in parallel relationships, the system reliability is higher than any of the individual component reliabilities.

EXAMPLE 3 *Series and Parallel Combinations*

Let's now consider a system combining series and parallel relationships, as we join together the previous two subsystems for a controlled propulsion ignition system (Figure 6.10). We examine each subsystem. Subsystem 1 (the communication system) is in a parallel relationship, and we found its reliability to be 0.998. Subsystem 2 (the propulsion system) is in a series relationship, and we found its system reliability to be 0.8208. These two systems are in a series relationship, so the reliability for the entire system is the product of the two subsystem reliabilities:

$$R_s(t) = R_{s_1}(t) \cdot R_{s_2}(t) = (0.998)(0.8208) = 0.8192$$

Figure 6.10
A NASA-controlled space shuttle propulsion ignition system



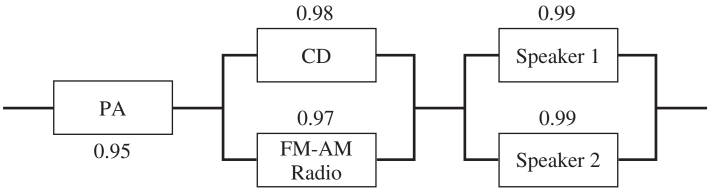
© Cengage Learning

6.2 Problems

1. Consider a stereo with CD player, FM–AM radio tuner, speakers (dual), and power amplifier (PA) components, as displayed with the reliabilities shown in Figure 6.11. Determine the system's reliability. What assumptions are required in your model?

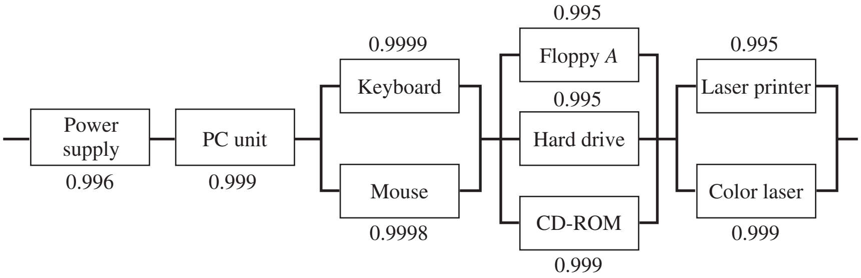
Figure 6.11

Reliability of stereo components



© Cengage Learning

2. Consider a personal computer with each item's reliability as shown in Figure 6.12. Determine the reliability of the computer system. What assumptions are required?



© Cengage Learning

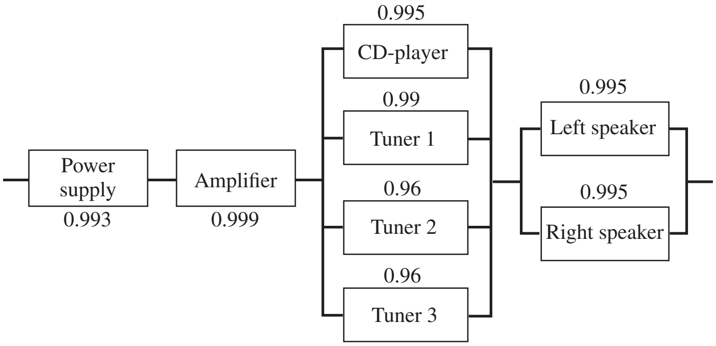
Figure 6.12

Personal computer reliabilities

3. Consider a more advanced stereo system with component reliabilities as displayed in Figure 6.13. Determine the system's reliability. What assumptions are required?

Figure 6.13

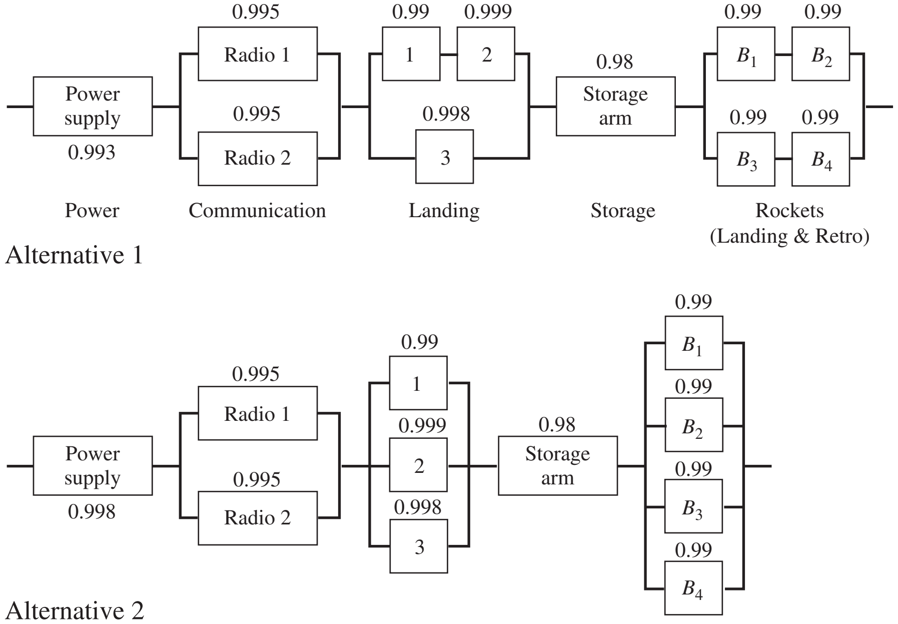
Advanced stereo system



© Cengage Learning

6.2 PROJECT

1. Two alternative designs are submitted for a landing module to enable the transport of astronauts to the surface of Mars. The mission is to land safely on Mars, collect several hundred pounds of samples from the planet's surface, and then return to the shuttle in its orbit around Mars. The alternative designs are displayed together with their reliabilities in Figure 6.14. Which design would you recommend to NASA? What assumptions are required? Are the assumptions reasonable?



© Cengage Learning

Figure 6.14
Alternative designs for the Mars module

6.3 Linear Regression

In Chapter 3 we discussed various criteria for fitting models to collected data. In particular, the least-squares criterion that minimizes the sum of the squared deviations was presented. We showed that the formulation of minimizing the sum of the squared deviations is an optimization problem. Until now, we have considered only a single observation y_i for each value of the independent variable x_i. However, what if we have multiple observations? In this section we explore a statistical methodology for minimizing the sum of the squared deviations, called **linear regression**. Our objectives are

1. To illustrate the basic linear regression model and its assumptions.
2. To define and interpret the statistic R^2.

3. To illustrate a graphical interpretation for the fit of the linear regression model by examining and interpreting the residual scatterplots.

We introduce only basic concepts and interpretations here. More in-depth studies of the subject of linear regression are given in advanced statistics courses.

The Linear Regression Model

The basic linear regression model is defined by

$$y_i = ax_i + b \quad \text{for } i = 1, 2, \ldots, m \text{ data points} \tag{6.4}$$

In Section 3.3 we derived the *normal equations*

$$\begin{aligned} a\sum_{i=1}^{m} x_i^2 + b\sum_{i=1}^{m} x_i &= \sum_{i=1}^{m} x_i y_i \\ a\sum_{i=1}^{m} x_i + mb &= \sum_{i=1}^{m} y_i \end{aligned} \tag{6.5}$$

and solved them to obtain the slope a and y-intercept b for the least-squares best-fitting line:

$$a = \frac{m\sum x_i y_i - \sum x_i \sum y_i}{m\sum x_i^2 - \left(\sum x_i\right)^2}, \text{ the slope} \tag{6.6}$$

and

$$b = \frac{\sum x_i^2 \sum y_i - \sum x_i y_i \sum x_i}{m\sum x_i^2 - \left(\sum x_i\right)^2}, \text{ the intercept} \tag{6.7}$$

We now add several additional equations to aid in the statistical analysis of the basic model (6.4).

The first of these is the **error sum of squares** given by

$$\text{SSE} = \sum_{i=1}^{m} [y_i - (ax_i + b)]^2 \tag{6.8}$$

which reflects variation about the regression line. The second concept is the **total corrected sum of squares** of y defined by

$$\text{SST} = \sum_{i=1}^{m} (y_i - \bar{y})^2 \tag{6.9}$$

where $\bar{y}$ is the average of the y values for the data points (x_i, y_i), $i = 1, \ldots, m$. (The number $\bar{y}$ is also the average value of the linear regression line $y = ax + b$ over the range of data.) Equations (6.8) and (6.9) then produce the **regression sum of squares** given by the equation

$$\text{SSR} = \text{SST} - \text{SSE} \tag{6.10}$$

The quantity SSR reflects the amount of variation in the y values explained by the linear regression line $y = ax + b$ when compared with the variation in the y values about the line $y = \bar{y}$.

From Equation (6.10), SST is always at least as large as SSE. This fact prompts the following definition of the **coefficient of determination** R^2, which is a measure of fit for the regression line.

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \tag{6.11}$$

The number R^2 expresses the proportion of the total variation in the y variable of the actual data (when compared with the line $y = \bar{y}$) that can be accounted for by the straight-line model, whose values are given by $ax + b$ (the predicted values) and calculated in terms of the x variable. If $R^2 = 0.81$, for instance, then 81% of the total variation of the y values (from the line $y = \bar{y}$) is accounted for by a linear relationship with the values of x. Thus, the closer the value of R^2 is to 1, the better the fit of the regression line model to the actual data. If $R^2 = 1$, then the data lie perfectly along the regression line. (Note that $R^2 \leq 1$ always holds.) The following are additional properties of R^2:

1. The value of R^2 does not depend on which of the two variables is labeled x and which is labeled y.
2. The value of R^2 is independent of the units of x and y.

Another indicator of the reasonableness of the fit is a plot of the **residuals** versus the independent variable. Recall that the residuals are the *errors* between the actual and predicted values:

$$r_i = y_i - f(x_i) = y_i - (ax_i - b) \tag{6.12}$$

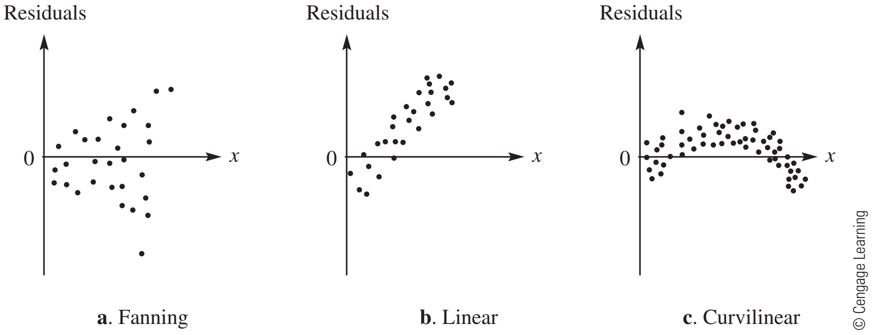
If we plot the residuals versus the independent variable, we obtain some valuable information about them:

1. The residuals should be randomly distributed and contained in a reasonably small band that is commensurate with the accuracy of the data.
2. An extremely large residual warrants further investigation of the associated data point to discover the cause of the large residual.
3. A pattern or trend in the residuals indicates that a forecastable effect remains to be modeled. The nature of the pattern often provides clues to how to refine the model, if a refinement is needed. These ideas are illustrated in Figure 6.15.

EXAMPLE 1 *Ponderosa Pines*

Recall the ponderosa pine data from Chapter 2, provided in Table 6.4. Figure 6.16 on page 237 shows a scatterplot of the data and suggests a trend. The plot is concave up and increasing, which could suggest a power function model (or an exponential model).

Problem Identification *Predict the number of board-feet as a function of the diameter of the ponderosa pine.*



a. Fanning **b**. Linear **c**. Curvilinear

Figure 6.15
Possible patterns of residuals plots

Table 6.4 Ponderosa pine data

Diameter (in.)	Board feet
36	192
28	113
28	88
41	294
19	28
32	123
22	51
38	252
25	56
17	16
31	141
20	32
25	86
19	21
39	231
33	187
17	22
37	205
23	57
39	265

© Cengage Learning

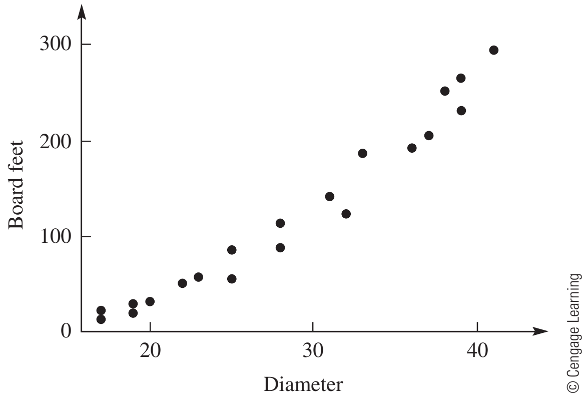
Assumptions We assume that the ponderosa pines are geometrically similar and have the shape of a right circular cylinder. This allows us to use the diameter as a characteristic dimension to predict the volume. We can reasonably assume that the height of a tree is proportional to its diameter.

Model Formulation Geometric similarity gives the proportionality

$$V \propto d^3 \tag{6.13}$$

Figure 6.16

Scatterplot of the ponderosa pine data



where d is the diameter of a tree (measured at ground level). If we further assume that the ponderosa pines have constant height (rather than assuming height is proportional to the diameter), then we obtain

$$V \propto d^2 \tag{6.14}$$

Assuming there is a constant volume associated with the underground root system would then suggest the following refinements of these proportionality models:

$$V = ad^3 + b \tag{6.15}$$

and

$$V = \alpha d^2 + \beta \tag{6.16}$$

Let's use linear regression to find the constant parameters in the four model types and then compare the results.

Model Solution The following solutions were obtained using a computer and applying linear regression on the four model types:

$$\begin{aligned} V &= 0.00431d^3 \\ V &= 0.00426d^3 + 2.08 \\ V &= 0.152d^2 \\ V &= 0.194d^2 - 45.7 \end{aligned}$$

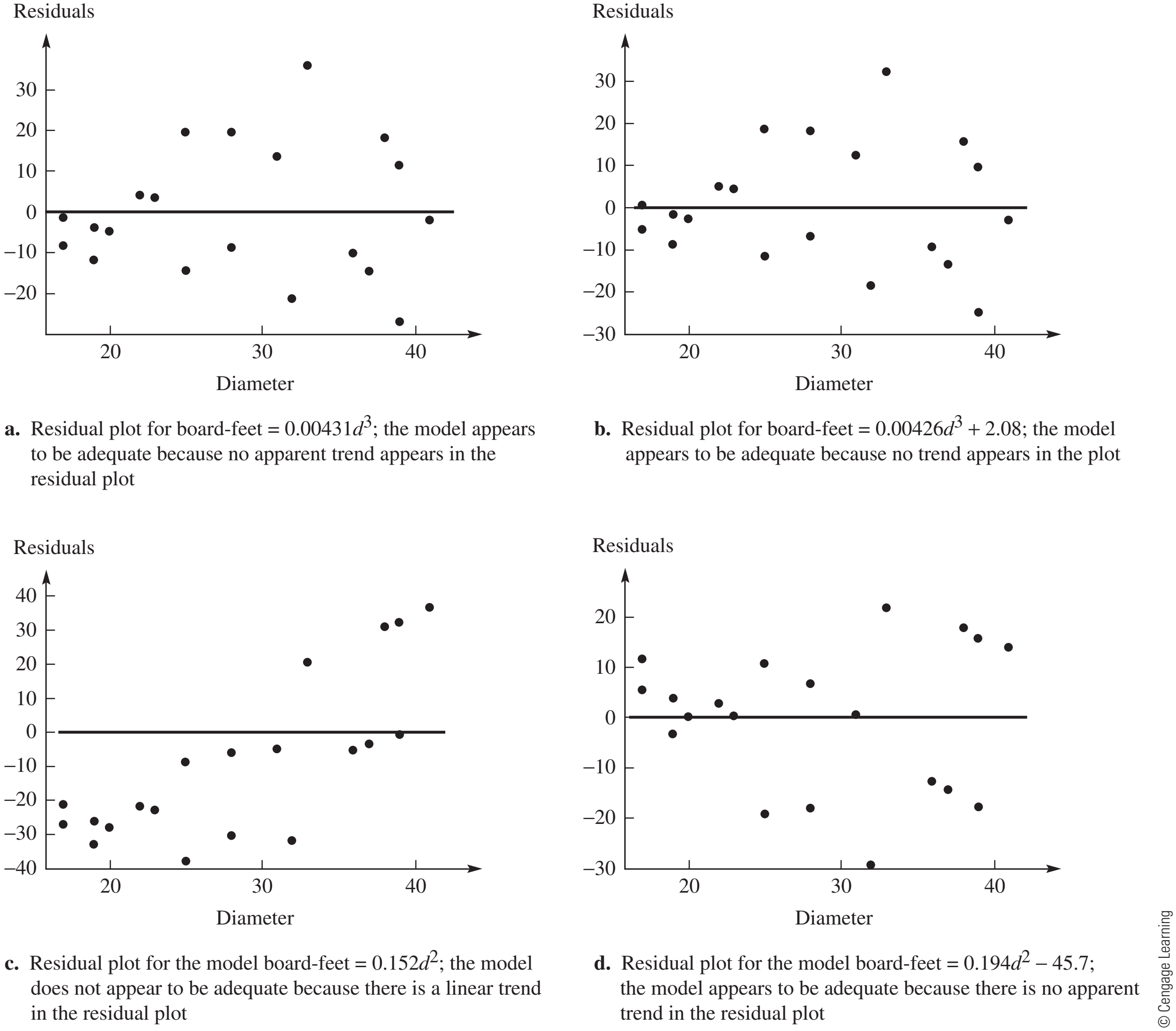
Table 6.5 displays the results of these regression models.

Notice that the R^2 values are all quite large (close to 1), which indicates a strong linear relationship. The residuals are calculated using Equation (6.12), and their plots are shown in Figure 6.17. (Recall that we are searching for a random distribution of the residuals having no apparent pattern.) Note that there is an apparent trend in the errors corresponding to the model $V = 0.152d^2$. We would probably *reject* (or *refine*) this model based on this plot, while accepting the other models that appear reasonable.

Table 6.5 Key information from regression models using the ponderosa pine data

Model	SSE	SSR	SST	R^2
$V = 0.00431d^3$	3,742	458,536	462,278	0.9919
$V = 0.00426d^3 + 2.08$	3,712	155,986	159,698	0.977
$V = 0.152d^2$	12,895	449,383	462,278	0.9721
$V = 0.194d^2 - 45.7$	3,910	155,788	159,698	0.976

© Cengage Learning



a. Residual plot for board-feet = $0.00431d^3$; the model appears to be adequate because no apparent trend appears in the residual plot

b. Residual plot for board-feet = $0.00426d^3 + 2.08$; the model appears to be adequate because no trend appears in the plot

c. Residual plot for the model board-feet = $0.152d^2$; the model does not appear to be adequate because there is a linear trend in the residual plot

d. Residual plot for the model board-feet = $0.194d^2 - 45.7$; the model appears to be adequate because there is no apparent trend in the residual plot

Figure 6.17
Residual plots for various models for board-feet = f(diameter) for the ponderosa pine data

EXAMPLE 2 *The Bass Fishing Derby Revisited*

Let's revisit the bass fishing problem from Section 2.3. We have collected much more data and now have 100 data points to use for fitting the model. These data are given in Table 6.6 and plotted in Figure 6.18. Based on the analysis in Section 2.3, we assume the following model type:

$$W = al^3 + b \tag{6.17}$$

where W is the weight of the fish and l is its length.

Table 6.6 Data for bass fish with weight (W) in oz and length (l) in in.

W	13	13	13	13	13	14	14	15	15	15
l	12	12.25	12	12.25	14.875	12	12	12.125	12.125	12.25
W	15	15	15	16	16	16	16	16	16	16
l	12	12.5	12.25	12.675	12.5	12.75	12.75	12	12.75	12.25
W	16	16	16	16	16	17	17	17	17	17
l	12	13	12.5	12.5	12.25	12.675	12.25	12.75	12.75	13.125
W	17	17	17	18	18	18	18	18	18	18
l	15.25	12.5	13.5	12.5	13	13.125	13	13.375	16.675	13
W	18	19	19	19	19	19	19	20	20	20
l	13.375	13.25	13.25	13.5	13.5	13.5	13	13.75	13.125	13.75
W	20	20	20	20	20	20	21	21	21	22
l	13.5	13.75	13.5	13.75	17	14.5	13.75	13.5	13.25	13.765
W	22	22	23	23	23	24	24	24	24	24
l	14	14	14.25	14.375	14	14.75	13.5	13.5	14.5	14
W	24	25	25	26	26	27	27	28	28	28
l	17	14.25	14.25	14.375	14.675	16.75	14.25	14.75	13	14.75
W	28	29	29	30	35	36	40	41	41	44
l	14.875	14.5	13.125	14.5	12.5	15.75	16.25	17.375	14.5	13.25
W	45	46	47	47	48	49	53	56	62	78
l	17.25	17	18	16.5	18	17.5	18	18.375	19.25	20

© Cengage Learning

The linear regression solution is

$$W = 0.008l^3 + 0.95 \tag{6.18}$$

and the analysis of variance is illustrated in Table 6.7.

The R^2 value for our model is 0.735, which is reasonably close to 1, considering the nature of the behavior being modeled. The residuals are found from Equation (6.12) and plotted in Figure 6.19. We see no apparent trend in the plot, so there is no suggestion on how to refine the model for possible improvement. ■ ■ ■