14

Linear Regression

CHAPTER OBJECTIVES

The primary objective of this chapter is to introduce you to how least-squares regression can be used to fit a straight line to measured data. Specific objectives and topics covered are

- Familiarizing yourself with some basic descriptive statistics and the normal distribution.
- Knowing how to compute the slope and intercept of a best-fit straight line with linear regression.
- Knowing how to generate random numbers with MATLAB and how they can be employed for Monte Carlo simulations.
- Knowing how to compute and understand the meaning of the coefficient of determination and the standard error of the estimate.
- Understanding how to use transformations to linearize nonlinear equations so that they can be fit with linear regression.
- Knowing how to implement linear regression with MATLAB.

YOU'VE GOT A PROBLEM

In Chap. 1, we noted that a free-falling object such as a bungee jumper is subject to the upward force of air resistance. As a first approximation, we assumed that this force was proportional to the square of velocity as in

$$F_U = c_d v^2 \tag{14.1}$$

where F_U = the upward force of air resistance [N = kg m/s^2], c_d = a drag coefficient (kg/m), and v = velocity [m/s].

Expressions such as Eq. (14.1) come from the field of fluid mechanics. Although such relationships derive in part from theory, experiments play a critical role in their formulation. One such experiment is depicted in Fig. 14.1. An individual is suspended in a wind

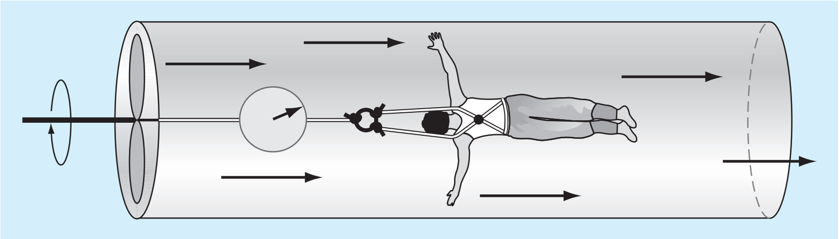


FIGURE 14.1
Wind tunnel experiment to measure how the force of air resistance depends on velocity.

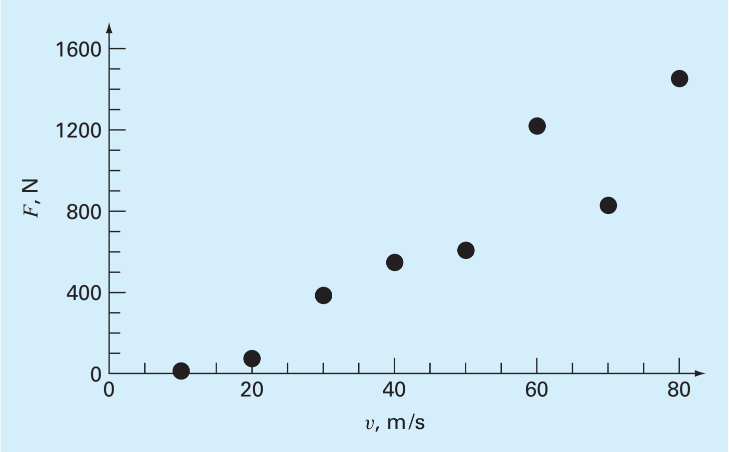


FIGURE 14.2
Plot of force versus wind velocity for an object suspended in a wind tunnel.

TABLE 14.1 Experimental data for force (N) and velocity (m/s) from a wind tunnel experiment.

v, m/s	10	20	30	40	50	60	70	80
F, N	25	70	380	550	610	1220	830	1450

tunnel (any volunteers?) and the force measured for various levels of wind velocity. The result might be as listed in Table 14.1.

The relationship can be visualized by plotting force versus velocity. As in Fig. 14.2, several features of the relationship bear mention. First, the points indicate that the force increases as velocity increases. Second, the points do not increase smoothly, but exhibit rather significant scatter, particularly at the higher velocities. Finally, although it may not be obvious, the relationship between force and velocity may not be linear. This conclusion becomes more apparent if we assume that force is zero for zero velocity.

In Chaps. 14 and 15, we will explore how to fit a “best” line or curve to such data. In so doing, we will illustrate how relationships like Eq. (14.1) arise from experimental data.

14.1 STATISTICS REVIEW

Before describing least-squares regression, we will first review some basic concepts from the field of statistics. These include the mean, standard deviation, residual sum of the squares, and the normal distribution. In addition, we describe how simple descriptive statistics and distributions can be generated in MATLAB. If you are familiar with these subjects, feel free to skip the following pages and proceed directly to Section 14.2. If you are unfamiliar with these concepts or are in need of a review, the following material is designed as a brief introduction.

14.1.1 Descriptive Statistics

Suppose that in the course of an engineering study, several measurements were made of a particular quantity. For example, Table 14.2 contains 24 readings of the coefficient of thermal expansion of a structural steel. Taken at face value, the data provide a limited amount of information—that is, that the values range from a minimum of 6.395 to a maximum of 6.775. Additional insight can be gained by summarizing the data in one or more well-chosen statistics that convey as much information as possible about specific characteristics of the data set. These descriptive statistics are most often selected to represent (1) the location of the center of the distribution of the data and (2) the degree of spread of the data set.

Measure of Location. The most common measure of central tendency is the arithmetic mean. The *arithmetic mean* ($\bar{y}$) of a sample is defined as the sum of the individual data points (y_i) divided by the number of points (n), or

$$\bar{y} = \frac{\sum y_i}{n} \tag{14.2}$$

where the summation (and all the succeeding summations in this section) is from $i = 1$ through n.

There are several alternatives to the arithmetic mean. The *median* is the midpoint of a group of data. It is calculated by first putting the data in ascending order. If the number of measurements is odd, the median is the middle value. If the number is even, it is the arithmetic mean of the two middle values. The median is sometimes called the *50th percentile.*

The *mode* is the value that occurs most frequently. The concept usually has direct utility only when dealing with discrete or coarsely rounded data. For continuous variables such as the data in Table 14.2, the concept is not very practical. For example, there are actually

TABLE 14.2 Measurements of the coefficient of thermal expansion of structural steel.

6.495	6.595	6.615	6.635	6.485	6.555
6.665	6.505	6.435	6.625	6.715	6.655
6.755	6.625	6.715	6.575	6.655	6.605
6.565	6.515	6.555	6.395	6.775	6.685

four modes for these data: 6.555, 6.625, 6.655, and 6.715, which all occur twice. If the numbers had not been rounded to 3 decimal digits, it would be unlikely that any of the values would even have repeated twice. However, if continuous data are grouped into equispaced intervals, it can be an informative statistic. We will return to the mode when we describe histograms later in this section.

Measures of Spread. The simplest measure of spread is the *range,* the difference between the largest and the smallest value. Although it is certainly easy to determine, it is not considered a very reliable measure because it is highly sensitive to the sample size and is very sensitive to extreme values.

The most common measure of spread for a sample is the *standard deviation* (s_y) about the mean:

$$s_y = \sqrt{\frac{S_t}{n-1}} \tag{14.3}$$

where S_t is the total sum of the squares of the residuals between the data points and the mean, or

$$S_t = \sum (y_i - \bar{y})^2 \tag{14.4}$$

Thus, if the individual measurements are spread out widely around the mean, S_t (and, consequently, s_y) will be large. If they are grouped tightly, the standard deviation will be small. The spread can also be represented by the square of the standard deviation, which is called the *variance:*

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} \tag{14.5}$$

Note that the denominator in both Eqs. (14.3) and (14.5) is $n - 1$. The quantity $n - 1$ is referred to as the *degrees of freedom.* Hence S_t and s_y are said to be based on $n - 1$ degrees of freedom. This nomenclature derives from the fact that the sum of the quantities upon which S_t is based (i.e., $\bar{y} - y_1, \bar{y} - y_2, \ldots, \bar{y} - y_n$) is zero. Consequently, if $\bar{y}$ is known and $n - 1$ of the values are specified, the remaining value is fixed. Thus, only $n - 1$ of the values are said to be freely determined. Another justification for dividing by $n - 1$ is the fact that there is no such thing as the spread of a single data point. For the case where $n = 1$, Eqs. (14.3) and (14.5) yield a meaningless result of infinity.

We should note that an alternative, more convenient formula is available to compute the variance:

$$s_y^2 = \frac{\sum y_i^2 - \left(\sum y_i\right)^2 / n}{n-1} \tag{14.6}$$

This version does not require precomputation of $\bar{y}$ and yields an identical result as Eq. (14.5).

A final statistic that has utility in quantifying the spread of data is the coefficient of variation (c.v.). This statistic is the ratio of the standard deviation to the mean. As such, it provides a normalized measure of the spread. It is often multiplied by 100 so that it can be expressed in the form of a percent:

$$\text{c.v.} = \frac{s_y}{\bar{y}} \times 100\% \tag{14.7}$$

EXAMPLE 14.1 Simple Statistics of a Sample

Problem Statement. Compute the mean, median, variance, standard deviation, and coefficient of variation for the data in Table 14.2.

Solution. The data can be assembled in tabular form and the necessary sums computed as in Table 14.3.

The mean can be computed as [Eq. (14.2)],

$$\bar{y} = \frac{158.4}{24} = 6.6$$

Because there are an even number of values, the median is computed as the arithmetic mean of the middle two values: $(6.605 + 6.615)/2 = 6.61$.

As in Table 14.3, the sum of the squares of the residuals is 0.217000, which can be used to compute the standard deviation [Eq. (14.3)]:

$$s_y = \sqrt{\frac{0.217000}{24 - 1}} = 0.097133$$

TABLE 14.3 Data and summations for computing simple descriptive statistics for the coefficients of thermal expansion from Table 14.2.

i	y_i	$(y_i - \bar{y})^2$	y_i^2
1	6.395	0.04203	40.896
2	6.435	0.02723	41.409
3	6.485	0.01323	42.055
4	6.495	0.01103	42.185
5	6.505	0.00903	42.315
6	6.515	0.00723	42.445
7	6.555	0.00203	42.968
8	6.555	0.00203	42.968
9	6.565	0.00123	43.099
10	6.575	0.00063	43.231
11	6.595	0.00003	43.494
12	6.605	0.00002	43.626
13	6.615	0.00022	43.758
14	6.625	0.00062	43.891
15	6.625	0.00062	43.891
16	6.635	0.00122	44.023
17	6.655	0.00302	44.289
18	6.655	0.00302	44.289
19	6.665	0.00422	44.422
20	6.685	0.00722	44.689
21	6.715	0.01322	45.091
22	6.715	0.01322	45.091
23	6.755	0.02402	45.630
24	6.775	0.03062	45.901
$\sum$	158.400	0.21700	1045.657

the variance [Eq. (14.5)]:

$$s_y^2 = (0.097133)^2 = 0.009435$$

and the coefficient of variation [Eq. (14.7)]:

$$\text{c.v.} = \frac{0.097133}{6.6} \times 100\% = 1.47\%$$

The validity of Eq. (14.6) can also be verified by computing

$$s_y^2 = \frac{1045.657 - (158.400)^2/24}{24 - 1} = 0.009435$$

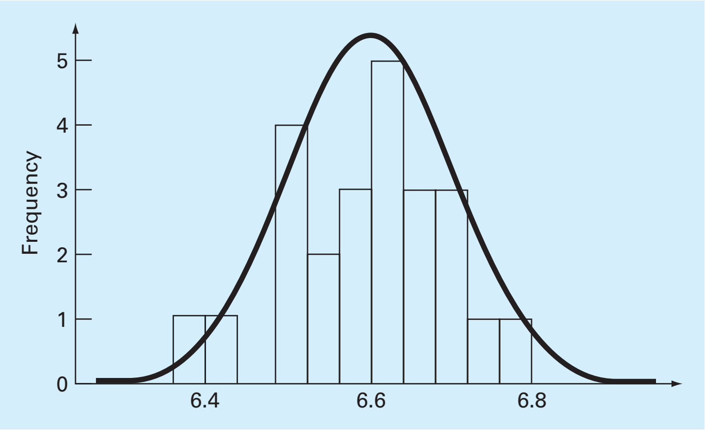
14.1.2 The Normal Distribution

Another characteristic that bears on the present discussion is the data distribution—that is, the shape with which the data are spread around the mean. A histogram provides a simple visual representation of the distribution. A *histogram* is constructed by sorting the measurements into intervals, or *bins*. The units of measurement are plotted on the abscissa and the frequency of occurrence of each interval is plotted on the ordinate.

As an example, a histogram can be created for the data from Table 14.2. The result (Fig. 14.3) suggests that most of the data are grouped close to the mean value of 6.6. Notice also, that now that we have grouped the data, we can see that the bin with the most values is from 6.6 to 6.64. Although we could say that the mode is the midpoint of this bin, 6.62, it is more common to report the most frequent range as the *modal class interval*.

If we have a very large set of data, the histogram often can be approximated by a smooth curve. The symmetric, bell-shaped curve superimposed on Fig. 14.3 is one such characteristic shape—the *normal distribution.* Given enough additional measurements, the histogram for this particular case could eventually approach the normal distribution.

FIGURE 14.3
A histogram used to depict the distribution of data. As the number of data points increases, the histogram often approaches the smooth, bell-shaped curve called the normal distribution.



The concepts of the mean, standard deviation, residual sum of the squares, and normal distribution all have great relevance to engineering and science. A very simple example is their use to quantify the confidence that can be ascribed to a particular measurement. If a quantity is normally distributed, the range defined by $\bar{y} - s_y$ to $\bar{y} + s_y$ will encompass approximately 68% of the total measurements. Similarly, the range defined by $\bar{y} - 2s_y$ to $\bar{y} + 2s_y$ will encompass approximately 95%.

For example, for the data in Table 14.2, we calculated in Example 14.1 that $\bar{y} = 6.6$ and $s_y = 0.097133$. Based on our analysis, we can tentatively make the statement that approximately 95% of the readings should fall between 6.405734 and 6.794266. Because it is so far outside these bounds, if someone told us that they had measured a value of 7.35, we would suspect that the measurement might be erroneous.

14.1.3 Descriptive Statistics in MATLAB

Standard MATLAB has several functions to compute descriptive statistics.[1] For example, the arithmetic mean is computed as `mean(x)`. If *x* is a vector, the function returns the mean of the vector's values. If it is a matrix, it returns a row vector containing the arithmetic mean of each column of *x*. The following is the result of using mean and the other statistical functions to analyze a column vector `s` that holds the data from Table 14.2:

```
>> format short g
>> mean(s),median(s),mode(s)

ans =
          6.6
ans =
         6.61
ans =
        6.555

>> min(s),max(s)

ans =
        6.395
ans =
        6.775

>> range=max(s)-min(s)

range =
         0.38

>> var(s),std(s)

ans =
    0.0094348

ans =
     0.097133
```

These results are consistent with those obtained previously in Example 14.1. Note that although there are four values that occur twice, the `mode` function only returns the first of the values: 6.555.

[1] MATLAB also offers a Statistics Toolbox that provides a wide range of common statistical tasks, from random number generation, to curve fitting, to design of experiments and statistical process control.

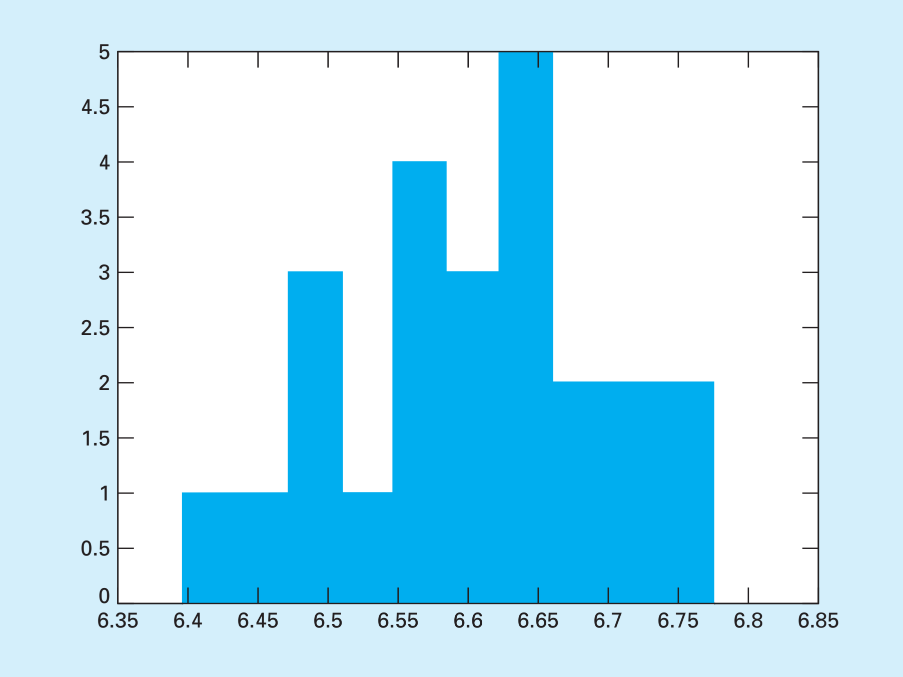


FIGURE 14.4
Histogram generated with the MATLAB `hist` function.

MATLAB can also be used to generate a histogram based on the `hist` function. The `hist` function has the syntax

```
[n, x] = hist(y, x)
```

where n = the number of elements in each bin, x = a vector specifying the midpoint of each bin, and y is the vector being analyzed. For the data from Table 14.2, the result is

```
>> [n,x] =hist(s)
n =
      1     1      3      1      4      3      5      2     2      2
x =
   6.414 6.452 6.49 6.528 6.566 6.604 6.642 6.68 6.718 6.756
```

The resulting histogram depicted in Fig. 14.4 is similar to the one we generated by hand in Fig. 14.3. Note that all the arguments and outputs with the exception of y are optional. For example, `hist(y)` without output arguments just produces a histogram bar plot with 10 bins determined automatically based on the range of values in y.

14.2 RANDOM NUMBERS AND SIMULATION

In this section, we will describe two MATLAB functions that can be used to produce a sequence of random numbers. The first (`rand`) generates numbers that are uniformly distributed, and the second (`randn`) generates numbers that have a normal distribution.

14.2.1 MATLAB Function: `rand`

This function generates a sequence of numbers that are uniformly distributed between 0 and 1. A simple representation of its syntax is

```
r = rand(m, n)
```

where `r` = an m-by-n matrix of random numbers. The following formula can then be used to generate a uniform distribution on another interval:

```
runiform = low + (up – low) * rand(m, n)
```

where `low` = the lower bound and `up` = the upper bound.

EXAMPLE 14.2 Generating Uniform Random Values of Drag

Problem Statement. If the initial velocity is zero, the downward velocity of the free-falling bungee jumper can be predicted with the following analytical solution (Eq. 1.9):

$$v = \sqrt{\frac{gm}{c_d}} \tanh\left(\sqrt{\frac{gc_d}{m}}\, t\right)$$

Suppose that $g = 9.81\text{m/s}^2$, and $m = 68.1$ kg, but c_d is not known precisely. For example, you might know that it varies uniformly between 0.225 and 0.275 (i.e., ±10% around a mean value of 0.25 kg/m). Use the `rand` function to generate 1000 random uniformly distributed values of c_d and then employ these values along with the analytical solution to compute the resulting distribution of velocities at $t = 4$ s.

Solution. Before generating the random numbers, we can first compute the mean velocity:

$$v_{\text{mean}} = \sqrt{\frac{9.81(68.1)}{0.25}} \tanh\left(\sqrt{\frac{9.81(0.25)}{68.1}}\, 4\right) = 33.1118 \frac{\text{m}}{\text{s}}$$

We can also generate the range:

$$v_{\text{low}} = \sqrt{\frac{9.81(68.1)}{0.275}} \tanh\left(\sqrt{\frac{9.81(0.275)}{68.1}}\, 4\right) = 32.6223 \frac{\text{m}}{\text{s}}$$

$$v_{\text{high}} = \sqrt{\frac{9.81(68.1)}{0.225}} \tanh\left(\sqrt{\frac{9.81(0.225)}{68.1}}\, 4\right) = 33.6198 \frac{\text{m}}{\text{s}}$$

Thus, we can see that the velocity varies by

$$\Delta v = \frac{33.6198 - 32.6223}{2(33.1118)} \times 100\% = 1.5063\%$$

The following script generates the random values for c_d, along with their mean, standard deviation, percent variation, and a histogram:

```
clc,format short g
n=1000;t=4;m=68.1;g=9.81;
cd=0.25;cdmin=cd-0.025,cdmax=cd+0.025
r=rand(n,1);
```

```
cdrand=cdmin+(cdmax-cdmin)*r;
meancd=mean(cdrand),stdcd=std(cdrand)
Deltacd=(max(cdrand)-min(cdrand))/meancd/2*100.
subplot(2,1,1)
hist(cdrand),title('(a) Distribution of drag')
xlabel('cd (kg/m)')
```

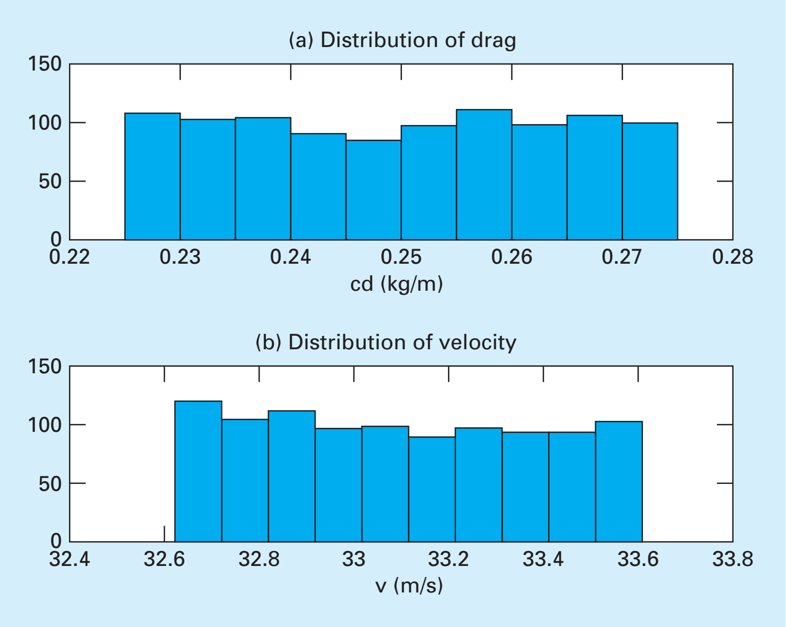
The results are

```
meancd =
      0.25018
stdcd =
     0.014528
Deltacd =
       9.9762
```

These results, as well as the histogram (Fig. 14.5*a*) indicate that `rand` has yielded 1000 uniformly distributed values with the desired mean value and range. The values can then be employed along with the analytical solution to compute the resulting distribution of velocities at $t = 4$ s.

```
vrand=sqrt(g*m./cdrand).*tanh(sqrt(g*cdrand/m)*t);
meanv=mean(vrand)
Deltav=(max(vrand)-min(vrand))/meanv/2*100.
subplot(2,1,2)
hist(vrand),title('(b) Distribution of velocity')
xlabel('v (m/s)')
```

FIGURE 14.5
Histograms of (*a*) uniformly distributed drag coefficients and (*b*) the resulting distribution of velocity.



The results are

```
meanv =
   33.1151
Deltav =
    1.5048
```

These results, as well as the histogram (Fig. 14.5*b*), closely conform to our hand calculations.

The foregoing example is formally referred to as a *Monte Carlo simulation*. The term, which is a reference to Monaco's Monte Carlo casino, was first used by physicists working on nuclear weapons projects in the 1940s. Although it yields intuitive results for this simple example, there are instances where such computer simulations yield surprising outcomes and provide insights that would otherwise be impossible to determine. The approach is feasible only because of the computer's ability to implement tedious, repetitive computations in an efficient manner.

14.2.2 MATLAB Function: `randn`

This function generates a sequence of numbers that are normally distributed with a mean of 0 and a standard deviation of 1. A simple representation of its syntax is

```
r = randn(m, n)
```

where `r` = an *m*-by-*n* matrix of random numbers. The following formula can then be used to generate a normal distribution with a different mean (`mn`) and standard deviation (`s`),

```
rnormal = mn + s * randn(m, n)
```

EXAMPLE 14.3 Generating Normally-Distributed Random Values of Drag

Problem Statement. Analyze the same case as in Example 14.2, but rather than employing a uniform distribution, generate normally-distributed drag coefficients with a mean of 0.25 and a standard deviation of 0.01443.

Solution. The following script generates the random values for c_d, along with their mean, standard deviation, coefficient of variation (expressed as a %), and a histogram:

```
clc,format short g
n=1000;t=4;m=68.1;g=9.81;
cd=0.25;
stdev=0.01443;
r=randn(n,1);
cdrand=cd+stdev*r;
meancd=mean(cdrand),stdevcd=std(cdrand)
cvcd=stdevcd/meancd*100.
subplot(2,1,1)
hist(cdrand),title('(a) Distribution of drag')
xlabel('cd (kg/m)')
```

The results are

```
meancd =
      0.24988
```

```
stdevcd =
     0.014465
cvcd =
       5.7887
```

These results, as well as the histogram (Fig. 14.6*a*) indicate that `randn` has yielded 1000 uniformly distributed values with the desired mean, standard deviation, and coefficient of variation. The values can then be employed along with the analytical solution to compute the resulting distribution of velocities at $t = 4$ s.

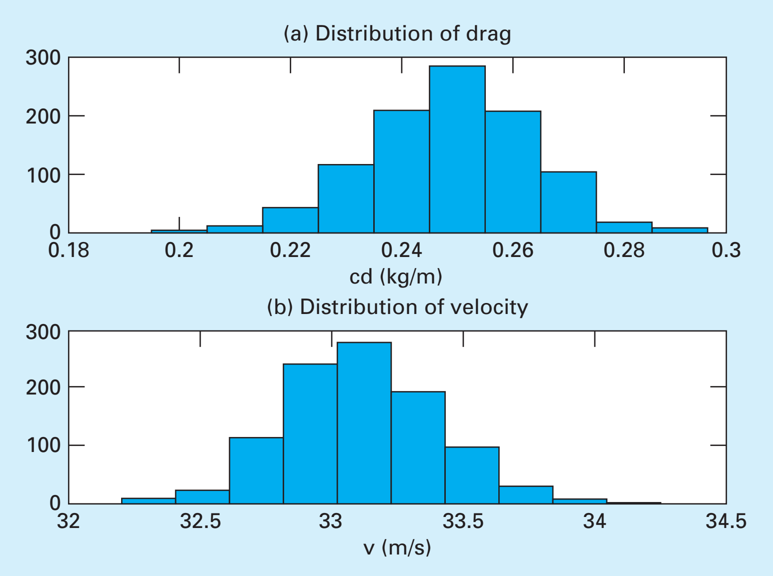
```
vrand=sqrt(g*m./cdrand).*tanh(sqrt(g*cdrand/m)*t);
meanv=mean(vrand),stdevv=std(vrand)
cvv=stdevv/meanv*100.
subplot(2,1,2)
hist(vrand),title('(b) Distribution of velocity')
xlabel('v (m/s)')
```

The results are

```
meanv =
       33.117
stdevv =
      0.28839
cvv =
       0.8708
```

These results, as well as the histogram (Fig. 14.6*b*), indicate that the velocities are also normally distributed with a mean that is close to the value that would be computed using the mean and the analytical solution. In addition, we compute the associated standard deviation which corresponds to a coefficient of variation of ±0.8708%.

FIGURE 14.6
Histograms of (a) normally-distributed drag coefficients and (b) the resulting distribution of velocity.



Although simple, the foregoing examples illustrate how random numbers can be easily generated within MATLAB. We will explore additional applications in the end-of-chapter problems.

14.3 LINEAR LEAST-SQUARES REGRESSION

Where substantial error is associated with data, the best curve-fitting strategy is to derive an approximating function that fits the shape or general trend of the data without necessarily matching the individual points. One approach to do this is to visually inspect the plotted data and then sketch a "best" line through the points. Although such "eyeball" approaches have commonsense appeal and are valid for "back-of-the-envelope" calculations, they are deficient because they are arbitrary. That is, unless the points define a perfect straight line (in which case, interpolation would be appropriate), different analysts would draw different lines.

To remove this subjectivity, some criterion must be devised to establish a basis for the fit. One way to do this is to derive a curve that minimizes the discrepancy between the data points and the curve. To do this, we must first quantify the discrepancy. The simplest example is fitting a straight line to a set of paired observations: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The mathematical expression for the straight line is

$$y = a_0 + a_1 x + e \tag{14.8}$$

where a_0 and a_1 are coefficients representing the intercept and the slope, respectively, and e is the error, or *residual,* between the model and the observations, which can be represented by rearranging Eq. (14.8) as

$$e = y - a_0 - a_1 x \tag{14.9}$$

Thus, the residual is the discrepancy between the true value of y and the approximate value, $a_0 + a_1 x$, predicted by the linear equation.

14.3.1 Criteria for a "Best" Fit

One strategy for fitting a "best" line through the data would be to minimize the sum of the residual errors for all the available data, as in

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) \tag{14.10}$$

where $n =$ total number of points. However, this is an inadequate criterion, as illustrated by Fig. 14.7*a*, which depicts the fit of a straight line to two points. Obviously, the best fit is the line connecting the points. However, any straight line passing through the midpoint of the connecting line (except a perfectly vertical line) results in a minimum value of Eq. (14.10) equal to zero because positive and negative errors cancel.

One way to remove the effect of the signs might be to minimize the sum of the absolute values of the discrepancies, as in

$$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - a_0 - a_1 x_i| \tag{14.11}$$

Figure 14.7*b* demonstrates why this criterion is also inadequate. For the four points shown, any straight line falling within the dashed lines will minimize the sum of the absolute values of the residuals. Thus, this criterion also does not yield a unique best fit.

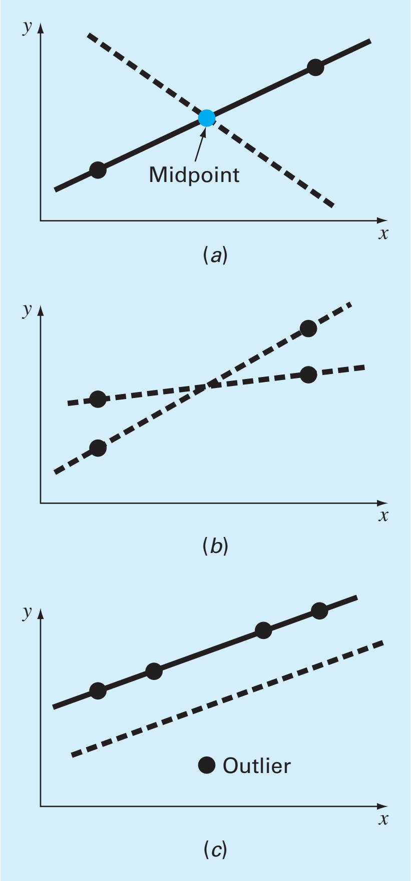


FIGURE 14.7
Examples of some criteria for "best fit" that are inadequate for regression: (*a*) minimizes the sum of the residuals, (*b*) minimizes the sum of the absolute values of the residuals, and (*c*) minimizes the maximum error of any individual point.

A third strategy for fitting a best line is the *minimax* criterion. In this technique, the line is chosen that minimizes the maximum distance that an individual point falls from the line. As depicted in Fig. 14.7*c*, this strategy is ill-suited for regression because it gives undue influence to an outlier—that is, a single point with a large error. It should be noted that the minimax principle is sometimes well-suited for fitting a simple function to a complicated function (Carnahan, Luther, and Wilkes, 1969).

A strategy that overcomes the shortcomings of the aforementioned approaches is to minimize the sum of the squares of the residuals:

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \tag{14.12}$$

This criterion, which is called *least squares,* has a number of advantages, including that it yields a unique line for a given set of data. Before discussing these properties, we will present a technique for determining the values of a_0 and a_1 that minimize Eq. (14.12).

14.3.2 Least-Squares Fit of a Straight Line

To determine values for a_0 and a_1, Eq. (14.12) is differentiated with respect to each unknown coefficient:

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i)$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum [(y_i - a_0 - a_1 x_i) x_i]$$

Note that we have simplified the summation symbols; unless otherwise indicated, all summations are from $i = 1$ to n. Setting these derivatives equal to zero will result in a minimum S_r. If this is done, the equations can be expressed as

$$0 = \sum y_i - \sum a_0 - \sum a_1 x_i$$

$$0 = \sum x_i y_i - \sum a_0 x_i - \sum a_1 x_i^2$$

Now, realizing that $\sum a_0 = n a_0$, we can express the equations as a set of two simultaneous linear equations with two unknowns (a_0 and a_1):

$$n \quad a_0 + \left(\sum x_i\right) a_1 = \sum y_i \tag{14.13}$$

$$\left(\sum x_i\right) a_0 + \left(\sum x_i^2\right) a_1 = \sum x_i y_i \tag{14.14}$$

These are called the *normal equations*. They can be solved simultaneously for

$$a_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2} \tag{14.15}$$

This result can then be used in conjunction with Eq. (14.13) to solve for

$$a_0 = \bar{y} - a_1 \bar{x} \tag{14.16}$$

where $\bar{y}$ and $\bar{x}$ are the means of y and x, respectively.

EXAMPLE 14.4 Linear Regression

Problem Statement. Fit a straight line to the values in Table 14.1.

Solution. In this application, force is the dependent variable (y) and velocity is the independent variable (x). The data can be set up in tabular form and the necessary sums computed as in Table 14.4.

TABLE 14.4 Data and summations needed to compute the best-fit line for the data from Table 14.1.

i	x_i	y_i	x_i^2	x_iy_i
1	10	25	100	250
2	20	70	400	1,400
3	30	380	900	11,400
4	40	550	1,600	22,000
5	50	610	2,500	30,500
6	60	1,220	3,600	73,200
7	70	830	4,900	58,100
8	80	1,450	6,400	116,000
$\sum$	360	5,135	20,400	312,850

The means can be computed as

$$\bar{x} = \frac{360}{8} = 45 \qquad \bar{y} = \frac{5{,}135}{8} = 641.875$$

The slope and the intercept can then be calculated with Eqs. (14.15) and (14.16) as

$$a_1 = \frac{8(312{,}850) - 360(5{,}135)}{8(20{,}400) - (360)^2} = 19.47024$$

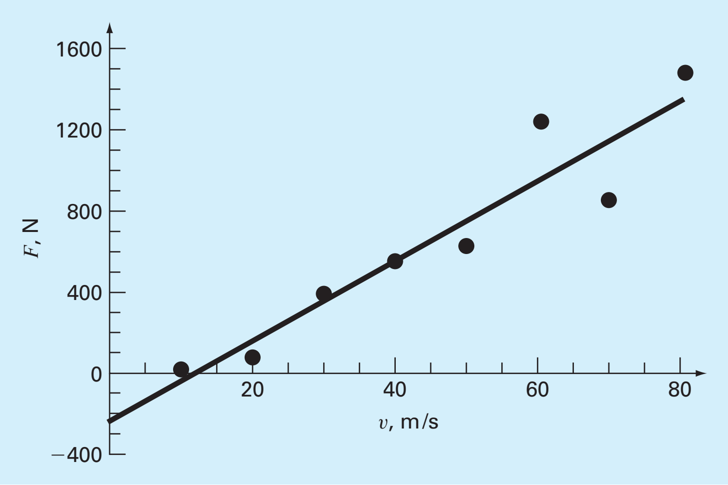
$$a_0 = 641.875 - 19.47024(45) = -234.2857$$

Using force and velocity in place of y and x, the least-squares fit is

$$F = -234.2857 + 19.47024v$$

The line, along with the data, is shown in Fig. 14.8.

FIGURE 14.8
Least-squares fit of a straight line to the data from Table 14.1



Notice that although the line fits the data well, the zero intercept means that the equation predicts physically unrealistic negative forces at low velocities. In Section 14.4, we will show how transformations can be employed to derive an alternative best-fit line that is more physically realistic.

14.3.3 Quantification of Error of Linear Regression

Any line other than the one computed in Example 14.4 results in a larger sum of the squares of the residuals. Thus, the line is unique and in terms of our chosen criterion is a "best" line through the points. A number of additional properties of this fit can be elucidated by examining more closely the way in which residuals were computed. Recall that the sum of the squares is defined as [Eq. (14.12)]

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \tag{14.17}$$

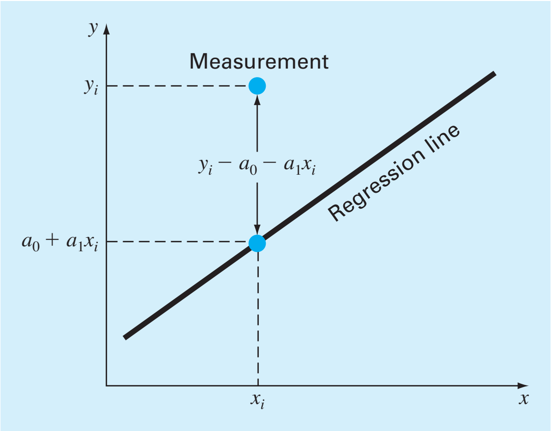
Notice the similarity between this equation and Eq. (14.4)

$$S_t = \sum (y_i - \bar{y})^2 \tag{14.18}$$

In Eq. (14.18), the square of the residual represented the square of the discrepancy between the data and a single estimate of the measure of central tendency—the mean. In Eq. (14.17), the square of the residual represents the square of the vertical distance between the data and another measure of central tendency—the straight line (Fig. 14.9).

The analogy can be extended further for cases where (1) the spread of the points around the line is of similar magnitude along the entire range of the data and (2) the distribution of these points about the line is normal. It can be demonstrated that if these criteria

FIGURE 14.9
The residual in linear regression represents the vertical distance between a data point and the straight line.



are met, least-squares regression will provide the best (i.e., the most likely) estimates of a_0 and a_1 (Draper and Smith, 1981). This is called the *maximum likelihood principle* in statistics. In addition, if these criteria are met, a "standard deviation" for the regression line can be determined as [compare with Eq. (14.3)]

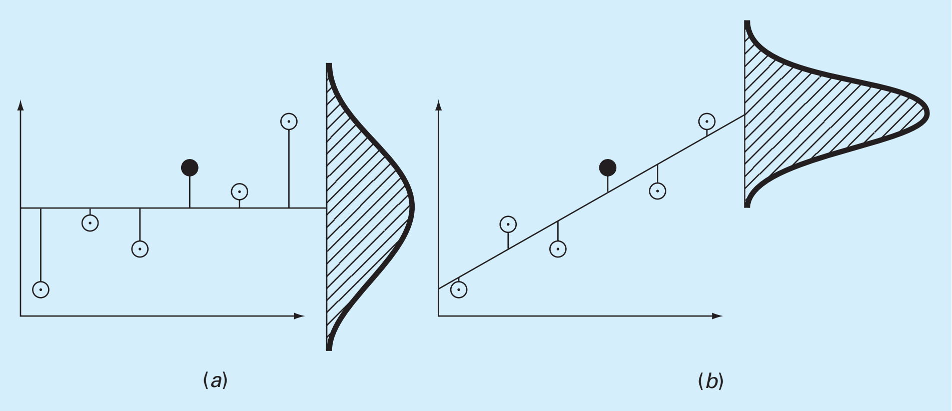
$$s_{y/x} = \sqrt{\frac{S_r}{n-2}} \tag{14.19}$$

where $s_{y/x}$ is called the *standard error of the estimate*. The subscript notation "y/x" designates that the error is for a predicted value of y corresponding to a particular value of x. Also, notice that we now divide by $n - 2$ because two data-derived estimates—a_0 and a_1—were used to compute S_r; thus, we have lost two degrees of freedom. As with our discussion of the standard deviation, another justification for dividing by $n - 2$ is that there is no such thing as the "spread of data" around a straight line connecting two points. Thus, for the case where $n = 2$, Eq. (14.19) yields a meaningless result of infinity.

Just as was the case with the standard deviation, the standard error of the estimate quantifies the spread of the data. However, $s_{y/x}$ quantifies the spread *around the regression line* as shown in Fig. 14.10*b* in contrast to the standard deviation s_y that quantified the spread *around the mean* (Fig. 14.10*a*).

These concepts can be used to quantify the "goodness" of our fit. This is particularly useful for comparison of several regressions (Fig. 14.11). To do this, we return to the original data and determine the total sum of the squares around the mean for the dependent variable (in our case, y). As was the case for Eq. (14.18), this quantity is designated S_t. This is the magnitude of the residual error associated with the dependent variable prior to regression. After performing the regression, we can compute S_r, the sum of the squares of the residuals around the regression line with Eq. (14.17). This characterizes the residual

FIGURE 14.10
Regression data showing (*a*) the spread of the data around the mean of the dependent variable and (*b*) the spread of the data around the best-fit line. The reduction in the spread in going from (*a*) to (*b*), as indicated by the bell-shaped curves at the right, represents the improvement due to linear regression.



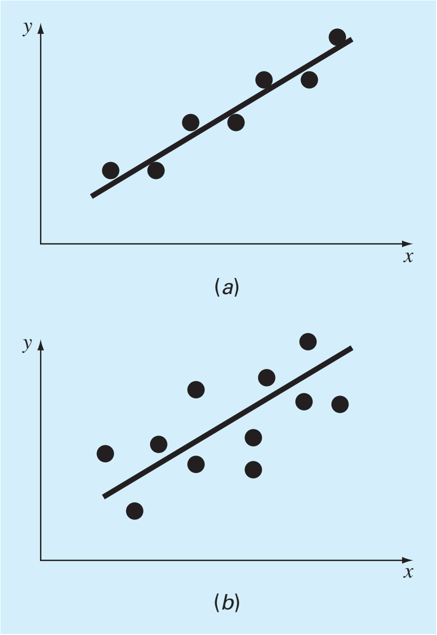


FIGURE 14.11
Examples of linear regression with (*a*) small and (*b*) large residual errors.

error that remains after the regression. It is, therefore, sometimes called the unexplained sum of the squares. The difference between the two quantities, $S_t - S_r$, quantifies the improvement or error reduction due to describing the data in terms of a straight line rather than as an average value. Because the magnitude of this quantity is scale-dependent, the difference is normalized to S_t to yield

$$r^2 = \frac{S_t - S_r}{S_t} \tag{14.20}$$

where r^2 is called the *coefficient of determination* and r is the *correlation coefficient* $(= \sqrt{r^2})$. For a perfect fit, $S_r = 0$ and $r^2 = 1$, signifying that the line explains 100% of the variability of the data. For $r^2 = 0$, $S_r = S_t$ and the fit represents no improvement. An alternative formulation for r that is more convenient for computer implementation is

$$r = \frac{n \sum(x_i y_i) - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n \sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n \sum y_i^2 - \left(\sum y_i\right)^2}} \tag{14.21}$$

EXAMPLE 14.5 Estimation of Errors for the Linear Least-Squares Fit

Problem Statement. Compute the total standard deviation, the standard error of the estimate, and the correlation coefficient for the fit in Example 14.4.

Solution. The data can be set up in tabular form and the necessary sums computed as in Table 14.5.

TABLE 14.5 Data and summations needed to compute the goodness-of-fit statistics for the data from Table 14.1.

i	x_i	y_i	$a_0 + a_1x_i$	$(y_i - \bar{y})^2$	$(y_i - a_0 - a_1x_i)^2$
1	10	25	−39.58	380,535	4,171
2	20	70	155.12	327,041	7,245
3	30	380	349.82	68,579	911
4	40	550	544.52	8,441	30
5	50	610	739.23	1,016	16,699
6	60	1,220	933.93	334,229	81,837
7	70	830	1,128.63	35,391	89,180
8	80	1,450	1,323.33	653,066	16,044
Σ	360	5,135		1,808,297	216,118

The standard deviation is [Eq. (14.3)]

$$s_y = \sqrt{\frac{1{,}808{,}297}{8-1}} = 508.26$$

and the standard error of the estimate is [Eq. (14.19)]

$$s_{y/x} = \sqrt{\frac{216{,}118}{8-2}} = 189.79$$

Thus, because $s_{y/x} < s_y$, the linear regression model has merit. The extent of the improvement is quantified by [Eq. (14.20)]

$$r^2 = \frac{1{,}808{,}297 - 216{,}118}{1{,}808{,}297} = 0.8805$$

or $r = \sqrt{0.8805} = 0.9383$. These results indicate that 88.05% of the original uncertainty has been explained by the linear model.

Before proceeding, a word of caution is in order. Although the coefficient of determination provides a handy measure of goodness-of-fit, you should be careful not to ascribe more meaning to it than is warranted. Just because r^2 is "close" to 1 does not mean that the fit is necessarily "good." For example, it is possible to obtain a relatively high value of r^2 when the underlying relationship between y and x is not even linear. Draper and Smith (1981) provide guidance and additional material regarding assessment of results for linear regression. In addition, at the minimum, you should always inspect a plot of the data along with your regression curve.

A nice example was developed by Anscombe (1973). As in Fig. 14.12, he came up with four data sets consisting of 11 data points each. Although their graphs are very different, all have the same best-fit equation, $y = 3 + 0.5x$, and the same coefficient of determination, $r^2 = 0.67$! This example dramatically illustrates why developing plots is so valuable.

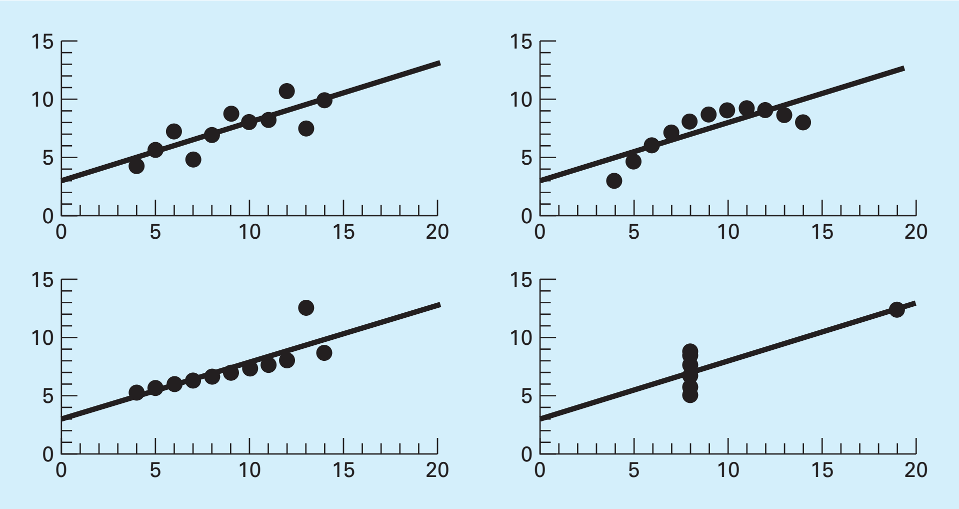


FIGURE 14.12
Anscombe's four data sets along with the best-fit line, $y = 3 + 0.5x$.

14.4 LINEARIZATION OF NONLINEAR RELATIONSHIPS

Linear regression provides a powerful technique for fitting a best line to data. However, it is predicated on the fact that the relationship between the dependent and independent variables is linear. This is not always the case, and the first step in any regression analysis should be to plot and visually inspect the data to ascertain whether a linear model applies. In some cases, techniques such as polynomial regression, which is described in Chap. 15, are appropriate. For others, transformations can be used to express the data in a form that is compatible with linear regression.

One example is the *exponential model:*

$$y = \alpha_1 e^{\beta_1 x} \tag{14.22}$$

where α_1 and β_1 are constants. This model is used in many fields of engineering and science to characterize quantities that increase (positive β_1) or decrease (negative β_1) at a rate that is directly proportional to their own magnitude. For example, population growth or radioactive decay can exhibit such behavior. As depicted in Fig. 14.13*a*, the equation represents a nonlinear relationship (for $\beta_1 \neq 0$) between y and x.

Another example of a nonlinear model is the simple *power equation:*

$$y = \alpha_2 x^{\beta_2} \tag{14.23}$$

where α_2 and β_2 are constant coefficients. This model has wide applicability in all fields of engineering and science. It is very frequently used to fit experimental data when the underlying model is not known. As depicted in Fig. 14.13*b*, the equation (for $\beta_2 \neq 0$) is nonlinear.

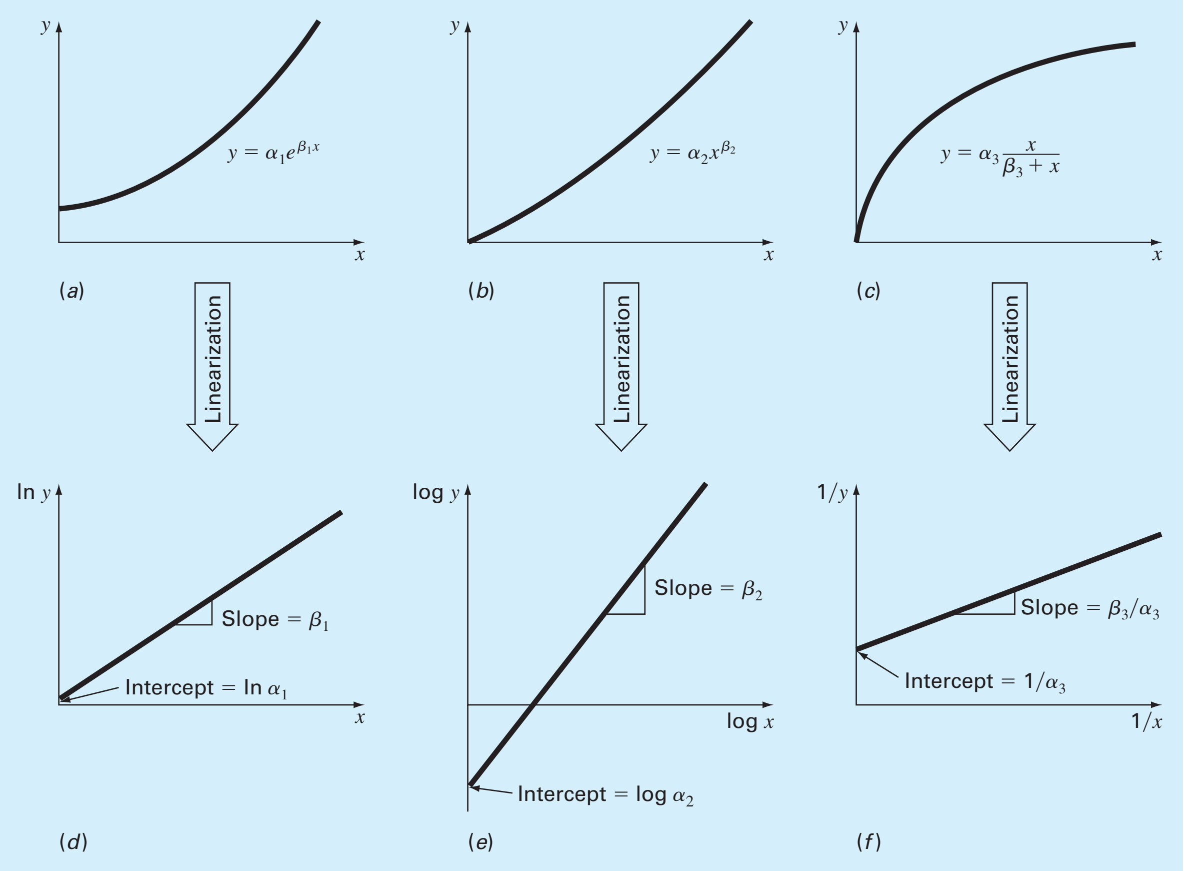

y
$y = \alpha_1 e^{\beta_1 x}$
x
(a)
y
$y = \alpha_2 x^{\beta_2}$
x
(b)
y
$y = \alpha_3 \frac{x}{\beta_3 + x}$
x
(c)
Linearization
Linearization
Linearization
ln y
Slope = β_1
Intercept = ln α_1
x
(d)
log y
Slope = β_2
log x
Intercept = log α_2
(e)
1/y
Slope = β_3/α_3
Intercept = $1/\alpha_3$
1/x
(f)

FIGURE 14.13
(*a*) The exponential equation, (*b*) the power equation, and (*c*) the saturation-growth-rate equation. Parts (*d*), (*e*), and (*f*) are linearized versions of these equations that result from simple transformations.

A third example of a nonlinear model is the *saturation-growth-rate equation:*

$$y = \alpha_3 \frac{x}{\beta_3 + x} \tag{14.24}$$

where α_3 and β_3 are constant coefficients. This model, which is particularly well-suited for characterizing population growth rate under limiting conditions, also represents a nonlinear relationship between y and x (Fig. 14.13*c*) that levels off, or "saturates," as x increases. It has many applications, particularly in biologically related areas of both engineering and science.

Nonlinear regression techniques are available to fit these equations to experimental data directly. However, a simpler alternative is to use mathematical manipulations to transform the equations into a linear form. Then linear regression can be employed to fit the equations to data.

For example, Eq. (14.22) can be linearized by taking its natural logarithm to yield

$$\ln y = \ln \alpha_1 + \beta_1 x \tag{14.25}$$

Thus, a plot of ln y versus x will yield a straight line with a slope of β_1 and an intercept of ln α_1 (Fig. 14.13*d*).

Equation (14.23) is linearized by taking its base-10 logarithm to give

$$\log y = \log \alpha_2 + \beta_2 \log x \tag{14.26}$$

Thus, a plot of log y versus log x will yield a straight line with a slope of β_2 and an intercept of log α_2 (Fig. 14.13*e*). Note that any base logarithm can be used to linearize this model. However, as done here, the base-10 logarithm is most commonly employed.

Equation (14.24) is linearized by inverting it to give

$$\frac{1}{y} = \frac{1}{\alpha_3} + \frac{\beta_3}{\alpha_3}\frac{1}{x} \tag{14.27}$$

Thus, a plot of $1/y$ versus $1/x$ will be linear, with a slope of β_3/α_3 and an intercept of $1/\alpha_3$ (Fig. 14.13*f*).

In their transformed forms, these models can be fit with linear regression to evaluate the constant coefficients. They can then be transformed back to their original state and used for predictive purposes. The following illustrates this procedure for the power model.

EXAMPLE 14.6 Fitting Data with the Power Equation

Problem Statement. Fit Eq. (14.23) to the data in Table 14.1 using a logarithmic transformation.

Solution. The data can be set up in tabular form and the necessary sums computed as in Table 14.6.

The means can be computed as

$$\bar{x} = \frac{12.606}{8} = 1.5757 \qquad \bar{y} = \frac{20.515}{8} = 2.5644$$

TABLE 14.6 Data and summations needed to fit the power model to the data from Table 14.1

i	x_i	y_i	$\log x_i$	$\log y_i$	$(\log x_i)^2$	$\log x_i \log y_i$
1	10	25	1.000	1.398	1.000	1.398
2	20	70	1.301	1.845	1.693	2.401
3	30	380	1.477	2.580	2.182	3.811
4	40	550	1.602	2.740	2.567	4.390
5	50	610	1.699	2.785	2.886	4.732
6	60	1220	1.778	3.086	3.162	5.488
7	70	830	1.845	2.919	3.404	5.386
8	80	1450	1.903	3.161	3.622	6.016
$\sum$			12.606	20.515	20.516	33.622

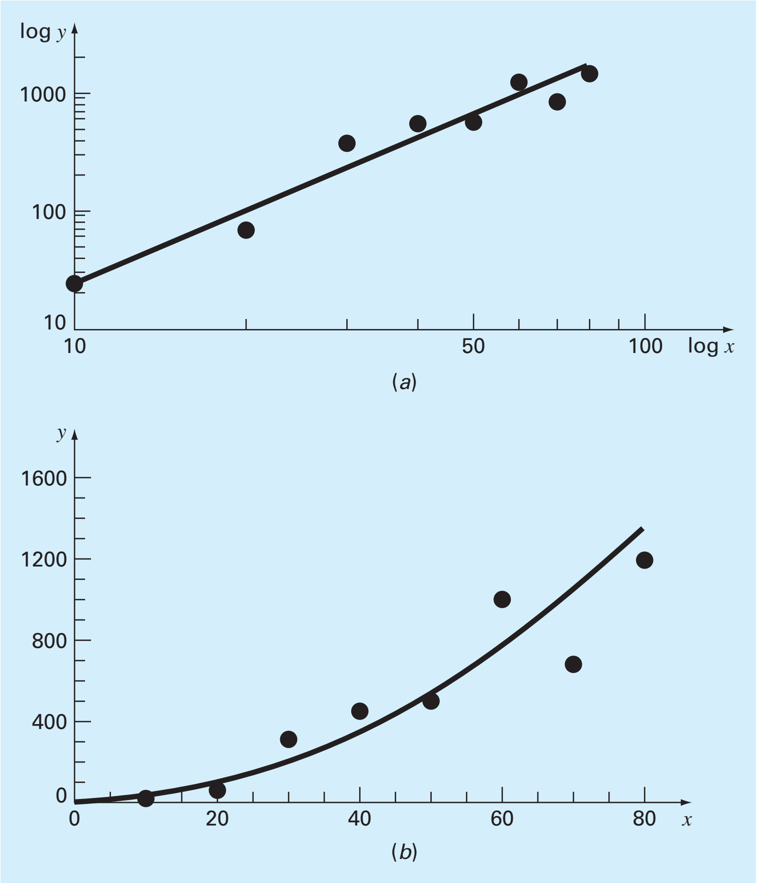


FIGURE 14.14
Least-squares fit of a power model to the data from Table 14.1. (a) The fit of the transformed data. (b) The power equation fit along with the data.

The slope and the intercept can then be calculated with Eqs. (14.15) and (14.16) as

$$a_1 = \frac{8(33.622) - 12.606(20.515)}{8(20.516) - (12.606)^2} = 1.9842$$

$$a_0 = 2.5644 - 1.9842(1.5757) = -0.5620$$

The least-squares fit is

$$\log y = -0.5620 + 1.9842 \log x$$

The fit, along with the data, is shown in Fig. 14.14a.

We can also display the fit using the untransformed coordinates. To do this, the coefficients of the power model are determined as $\alpha_2 = 10^{-0.5620} = 0.2741$ and $\beta_2 = 1.9842$. Using force and velocity in place of y and x, the least-squares fit is

$$F = 0.2741v^{1.9842}$$

This equation, along with the data, is shown in Fig. 14.14*b*.

The fits in Example 14.6 (Fig. 14.14) should be compared with the one obtained previously in Example 14.4 (Fig. 14.8) using linear regression on the untransformed data. Although both results would appear to be acceptable, the transformed result has the advantage that it does not yield negative force predictions at low velocities. Further, it is known from the discipline of fluid mechanics that the drag force on an object moving through a fluid is often well described by a model with velocity squared. Thus, knowledge from the field you are studying often has a large bearing on the choice of the appropriate model equation you use for curve fitting.

14.4.1 General Comments on Linear Regression

Before proceeding to curvilinear and multiple linear regression, we must emphasize the introductory nature of the foregoing material on linear regression. We have focused on the simple derivation and practical use of equations to fit data. You should be cognizant of the fact that there are theoretical aspects of regression that are of practical importance but are beyond the scope of this book. For example, some statistical assumptions that are inherent in the linear least-squares procedures are

1. Each x has a fixed value; it is not random and is known without error.
2. The y values are independent random variables and all have the same variance.
3. The y values for a given x must be normally distributed.

Such assumptions are relevant to the proper derivation and use of regression. For example, the first assumption means that (1) the x values must be error-free and (2) the regression of y versus x is not the same as x versus y. You are urged to consult other references such as Draper and Smith (1981) to appreciate aspects and nuances of regression that are beyond the scope of this book.

14.5 COMPUTER APPLICATIONS

Linear regression is so commonplace that it can be implemented on most pocket calculators. In this section, we will show how a simple M-file can be developed to determine the slope and intercept as well as to create a plot of the data and the best-fit line. We will also show how linear regression can be implemented with the built-in `polyfit` function.

14.5.1 MATLAB M-file: `linregr`

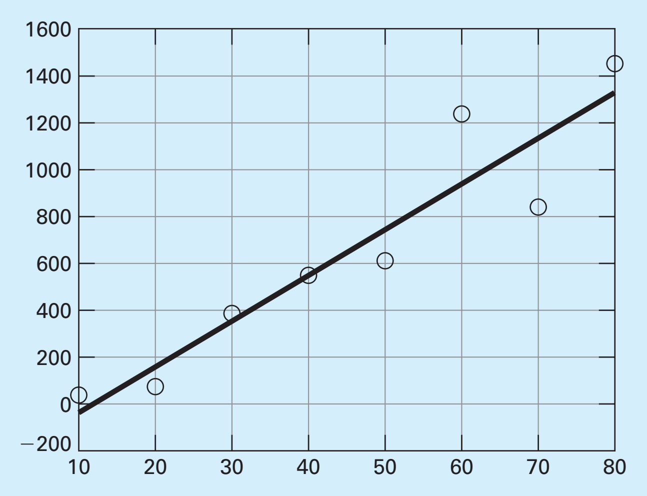
An algorithm for linear regression can be easily developed (Fig. 14.15). The required summations are readily computed with MATLAB's `sum` function. These are then used to compute the slope and the intercept with Eqs. (14.15) and (14.16). The routine displays the intercept and slope, the coefficient of determination, and a plot of the best-fit line along with the measurements.

A simple example of the use of this M-file would be to fit the force-velocity data analyzed in Example 14.4:

```
>> x = [10 20 30 40 50 60 70 80];
>> y = [25 70 380 550 610 1220 830 1450];
>> linregr(x,y)

r2 =
    0.8805

ans =
   19.4702 -234.2857
```



It can just as easily be used to fit the power model (Example 14.6) by applying the `log10` function to the data as in

```
>> linregr(log10(x),log10(y))

r2 =
    0.9481

ans =
    1.9842   -0.5620
```

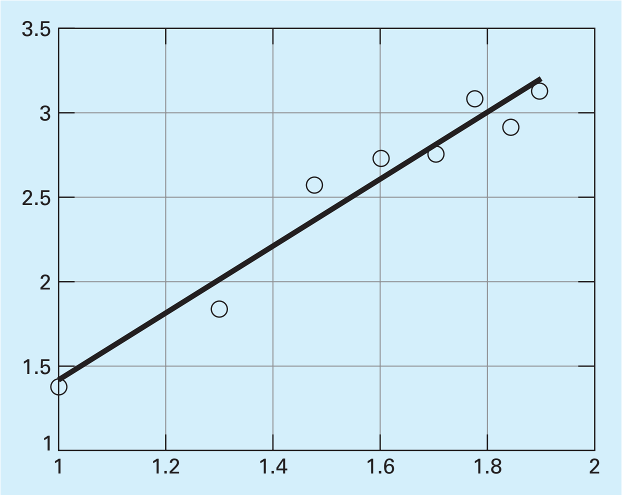


FIGURE 14.15
An M-file to implement linear regression.

```
function [a, r2] = linregr(x,y)
% linregr: linear regression curve fitting
%    [a, r2] = linregr(x,y):Least squares fit of straight
%             line to data by solving the normal equations

% input:
%   x = independent variable
%   y = dependent variable
% output:
%   a = vector of slope, a(1), and intercept, a(2)
%   r2 = coefficient of determination

n = length(x);
if length(y)~=n, error('x and y must be same length'); end
x = x(:); y = y(:);     % convert to column vectors
sx = sum(x); sy = sum(y);
sx2 = sum(x.*x); sxy = sum(x.*y); sy2 = sum(y.*y);
a(1) = (n*sxy-sx*sy)/(n*sx2-sx^2);
a(2) = sy/n-a(1)*sx/n;
r2 = ((n*sxy-sx*sy)/sqrt(n*sx2-sx^2)/sqrt(n*sy2-sy^2))^2;
% create plot of data and best fit line
xp = linspace(min(x),max(x),2);
yp = a(1)*xp+a(2);
plot(x,y,'o',xp,yp)
grid on
```

14.5.2 MATLAB Functions: `polyfit` and `polyval`

MATLAB has a built-in function `polyfit` that fits a least-squares *n*th-order polynomial to data. It can be applied as in

```
>> p = polyfit(x, y, n)
```

where `x` and `y` are the vectors of the independent and the dependent variables, respectively, and `n` = the order of the polynomial. The function returns a vector `p` containing the polynomial's coefficients. We should note that it represents the polynomial using decreasing powers of x as in the following representation:

$$f(x) = p_1 x^n + p_2 x^{n-1} + \cdots + p_n x + p_{n+1}$$

Because a straight line is a first-order polynomial, `polyfit(x,y,1)` will return the slope and the intercept of the best-fit straight line.

```
>> x = [10 20 30 40 50 60 70 80];
>> y = [25 70 380 550 610 1220 830 1450];
>> a = polyfit(x,y,1)

a =
   19.4702 -234.2857
```

Thus, the slope is 19.4702 and the intercept is −234.2857.

Another function, `polyval`, can then be used to compute a value using the coefficients. It has the general format:

```
>> y = polyval(p, x)
```

where `p` = the polynomial coefficients, and `y` = the best-fit value at `x`. For example,

```
>> y = polyval(a,45)

y =
  641.8750
```

14.6 CASE STUDY ENZYME KINETICS

Background. *Enzymes* act as catalysts to speed up the rate of chemical reactions in living cells. In most cases, they convert one chemical, the *substrate,* into another, the *product*. The *Michaelis-Menten* equation is commonly used to describe such reactions:

$$v = \frac{v_m[S]}{k_s + [S]} \tag{14.28}$$

where v = the initial reaction velocity, v_m = the maximum initial reaction velocity, $[S]$ = substrate concentration, and k_s = a half-saturation constant. As in Fig. 14.16, the equation describes a saturating relationship which levels off with increasing $[S]$. The graph also illustrates that the *half-saturation constant* corresponds to the substrate concentration at which the velocity is half the maximum.