CHAPTER 17

Least-Squares Regression

Where substantial error is associated with data, polynomial interpolation is inappropriate and may yield unsatisfactory results when used to predict intermediate values. Experimental data are often of this type. For example, Fig. 17.1*a* shows seven experimentally derived data points exhibiting significant variability. Visual inspection of these data suggests a positive relationship between y and x. That is, the overall trend indicates that higher values of y are associated with higher values of x. Now, if a sixth-order interpolating polynomial is fitted to these data (Fig. 17.1*b*), it will pass exactly through all of the points. However, because of the variability in these data, the curve oscillates widely in the interval between the points. In particular, the interpolated values at $x = 1.5$ and $x = 6.5$ appear to be well beyond the range suggested by these data.

A more appropriate strategy for such cases is to derive an approximating function that fits the shape or general trend of the data without necessarily matching the individual points. Figure 17.1*c* illustrates how a straight line can be used to generally characterize the trend of these data without passing through any particular point.

One way to determine the line in Fig. 17.1*c* 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. A technique for accomplishing this objective, called *least-squares regression,* will be discussed in the present chapter.

17.1 LINEAR REGRESSION

The simplest example of a least-squares approximation 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

$$y = a_0 + a_1 x + e \tag{17.1}$$

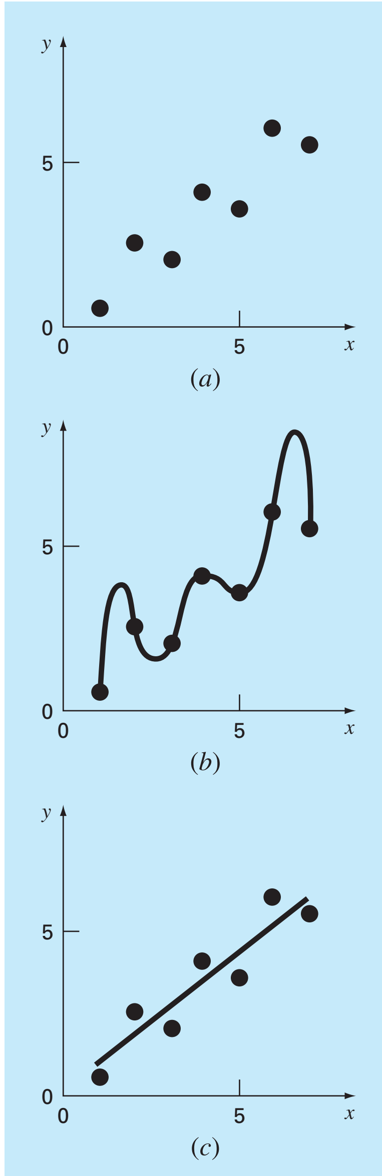


FIGURE 17.1
(*a*) Data exhibiting significant error. (*b*) Polynomial fit oscillating beyond the range of the data. (*c*) More satisfactory result using the least-squares fit.

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. (17.1) as

$$e = y - a_0 - a_1 x$$

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

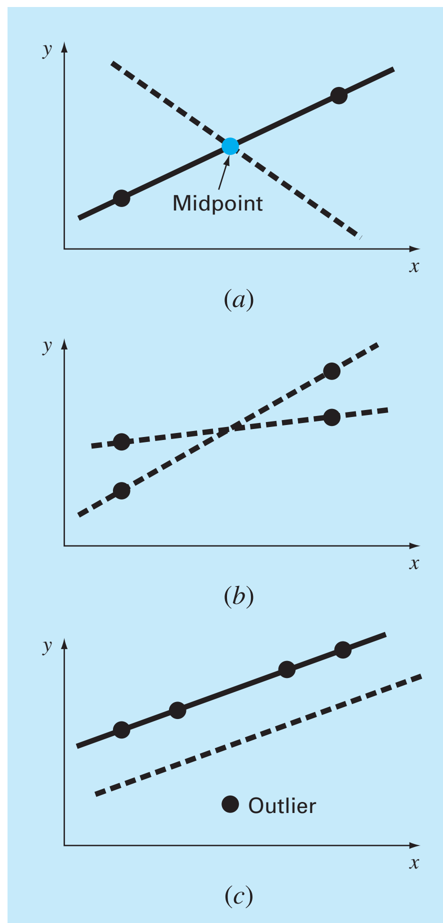
17.1.1 Criteria for a "Best" Fit

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

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) \tag{17.2}$$

where n = total number of points. However, this is an inadequate criterion, as illustrated by Fig. 17.2*a* which depicts the fit of a straight line to two points. Obviously, the best

FIGURE 17.2
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.



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. (17.2) equal to zero because the errors cancel.

Therefore, another logical criterion might be to minimize the sum of the absolute values of the discrepancies, as in

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

Figure 17.2*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. Thus, this criterion also does not yield a unique best fit.

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. 17.2*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 between the measured y and the y calculated with the linear model

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_{i,\text{measured}} - y_{i,\text{model}})^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \tag{17.3}$$

This criterion has a number of advantages, including the fact 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. (17.3).

17.1.2 Least-Squares Fit of a Straight Line

To determine values for a_0 and a_1, Eq. (17.3) is differentiated with respect to each 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

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

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

Now, realizing that $\Sigma a_0 = na_0$, we can express the equations as a set of two simultaneous linear equations with two unknowns (a_0 and a_1):

$$na_0 + \left(\sum x_i\right)a_1 = \sum y_i \tag{17.4}$$

$$\left(\sum x_i\right)a_0 + \left(\sum x_i^2\right)a_1 = \sum x_i y_i \tag{17.5}$$

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

$$a_1 = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} \tag{17.6}$$

This result can then be used in conjunction with Eq. (17.4) to solve for

$$a_0 = \bar{y} - a_1\bar{x} \tag{17.7}$$

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

EXAMPLE 17.1

Linear Regression

Problem Statement. Fit a straight line to the x and y values in the first two columns of Table 17.1.

Solution. The following quantities can be computed:

$$n = 7 \qquad \sum x_i y_i = 119.5 \qquad \sum x_i^2 = 140$$

$$\sum x_i = 28 \qquad \bar{x} = \frac{28}{7} = 4$$

$$\sum y_i = 24 \qquad \bar{y} = \frac{24}{7} = 3.428571$$

Using Eqs. (17.6) and (17.7),

$$a_1 = \frac{7(119.5) - 28(24)}{7(140) - (28)^2} = 0.8392857$$

$$a_0 = 3.428571 - 0.8392857(4) = 0.07142857$$

TABLE 17.1 Computations for an error analysis of the linear fit.

x_i	y_i	$(y_i - \bar{y})$	$(y_i - a_0 - a_1x_i)^2$
1	0.5	8.5765	0.1687
2	2.5	0.8622	0.5625
3	2.0	2.0408	0.3473
4	4.0	0.3265	0.3265
5	3.5	0.0051	0.5896
6	6.0	6.6122	0.7972
7	5.5	4.2908	0.1993
Σ	24.0	22.7143	2.9911

Therefore, the least-squares fit is

$$y = 0.07142857 + 0.8392857x$$

The line, along with the data, is shown in Fig. 17.1*c*.

17.1.3 Quantification of Error of Linear Regression

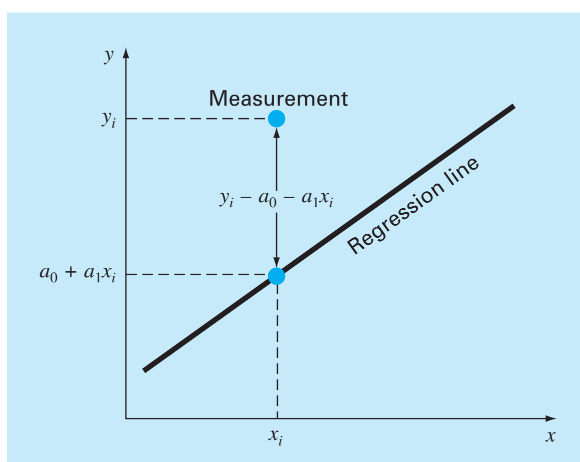
Any line other than the one computed in Example 17.1 results in a larger sum of the squares of the residuals. Thus, the line is unique and in terms of our chosen criterion is a "best" line through the points. A number of additional properties of this fit can be elucidated by examining more closely the way in which residuals were computed. Recall that the sum of the squares is defined as [Eq. (17.3)]

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \tag{17.8}$$

Notice the similarity between Eqs. (PT5.3) and (17.8). In the former case, 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. (17.8), 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. 17.3).

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 are met, least-squares regression will provide the best (that is, the most likely) estimates of a_0 and a_1 (Draper and Smith, 1981). This is called the *maximum likelihood*

FIGURE 17.3
The residual in linear regression represents the vertical distance between a data point and the straight line.



principle in statistics. In addition, if these criteria are met, a "standard deviation" for the regression line can be determined as [compare with Eq. (PT5.2)]

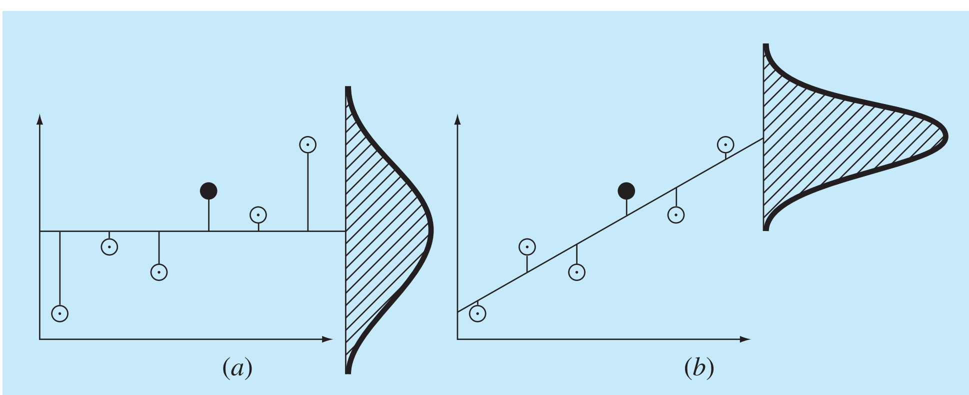
$$s_{y/x} = \sqrt{\frac{S_r}{n-2}} \tag{17.9}$$

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 in PT5.2.1, 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. (17.9) 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. 17.4*b* in contrast to the original standard deviation s_y that quantified the spread *around the mean* (Fig. 17.4*a*).

The above concepts can be used to quantify the "goodness" of our fit. This is particularly useful for comparison of several regressions (Fig. 17.5). 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. (PT5.3), 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. This characterizes the residual error that remains after the regression. It is, therefore, sometimes called the unexplained

FIGURE 17.4
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.



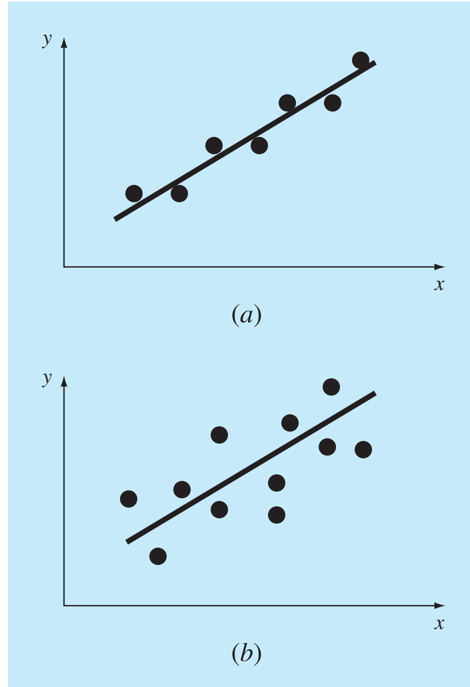


FIGURE 17.5
Examples of linear regression with (*a*) small and (*b*) large residual errors.

EXAMPLE 17.2 Estimation of Errors for the Linear Least-Squares Fit

Problem Statement. Compute the standard error of the estimate, for the data in Example 17.1.

Solution. The summations are performed and presented in Table 17.1.

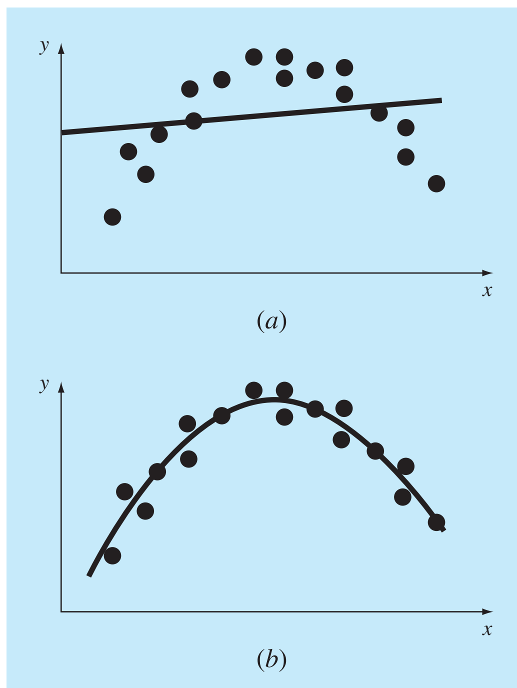
and the standard error of the estimate is [Eq. (17.9)]

$$s_{y/x} = \sqrt{\frac{2.9911}{7-2}} = 0.7735$$

17.1.5 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. For example, Fig. 17.8 shows some data that is obviously curvilinear. In some cases, techniques such as polynomial regression, which is described in Sec. 17.2, are appropriate. For others, transformations can be used to express the data in a form that is compatible with linear regression.

FIGURE 17.8
(*a*) Data that are ill-suited for linear least-squares regression. (*b*) Indication that a parabola is preferable.



One example is the *exponential model*

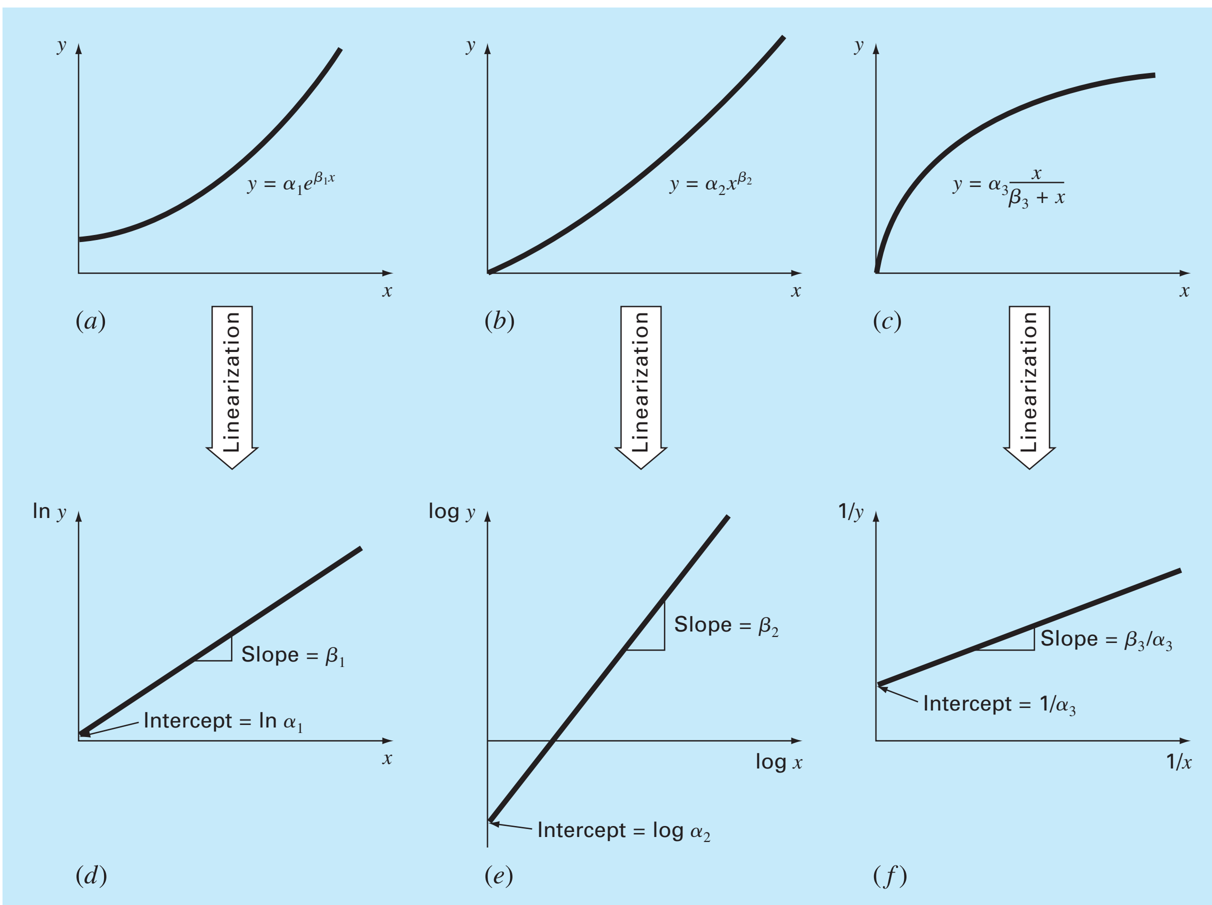
$$y = \alpha_1 e^{\beta_1 x} \tag{17.12}$$

where α_1 and β_1 are constants. This model is used in many fields of engineering 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. 17.9*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

$$y = \alpha_2 x^{\beta_2} \tag{17.13}$$

FIGURE 17.9
(*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.



where α_2 and β_2 are constant coefficients. This model has wide applicability in all fields of engineering. As depicted in Fig. 17.9*b*, the equation (for $\beta_2 \neq 0$ or 1) is nonlinear.

A third example of a nonlinear model is the saturation-growth-rate equation [recall Eq. (E17.3.1)]

$$y = \alpha_3 \frac{x}{\beta_3 + x} \tag{17.14}$$

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. 17.9*c*) that levels off, or "saturates," as x increases.

Nonlinear regression techniques are available to fit these equations to experimental data directly. (Note that we will discuss nonlinear regression in Sec. 17.5.) However, a simpler alternative is to use mathematical manipulations to transform the equations into a linear form. Then, simple linear regression can be employed to fit the equations to data.

For example, Eq. (17.12) can be linearized by taking its natural logarithm to yield

$$\ln y = \ln \alpha_1 + \beta_1 x \ln e$$

But because $\ln e = 1$,

$$\ln y = \ln \alpha_1 + \beta_1 x \tag{17.15}$$

Thus, a plot of ln y versus x will yield a straight line with a slope of β_1 and an intercept of $\ln \alpha_1$ (Fig. 17.9*d*).

Equation (17.13) is linearized by taking its base-10 logarithm to give

$$\log y = \beta_2 \log x + \log \alpha_2 \tag{17.16}$$

Thus, a plot of log y versus log x will yield a straight line with a slope of β_2 and an intercept of $\log \alpha_2$ (Fig. 17.9*e*).

Equation (17.14) is linearized by inverting it to give

$$\frac{1}{y} = \frac{\beta_3}{\alpha_3}\frac{1}{x} + \frac{1}{\alpha_3} \tag{17.17}$$

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. 17.9*f*).

In their transformed forms, these models can use linear regression to evaluate the constant coefficients. They could then be transformed back to their original state and used for predictive purposes. Example 17.4 illustrates this procedure for Eq. (17.13). In addition, Sec. 20.1 provides an engineering example of the same sort of computation.

EXAMPLE 17.4 Linearization of a Power Equation

Problem Statement. Fit Eq. (17.13) to the data in Table 17.3 using a logarithmic transformation of the data.

Solution. Figure 17.10*a* is a plot of the original data in its untransformed state. Figure 17.10*b* shows the plot of the transformed data. A linear regression of the log-transformed data yields the result

$$\log y = 1.75 \log x - 0.300$$

TABLE 17.3 Data to be fit to the power equation.

x	y	log x	log y
1	0.5	0	−0.301
2	1.7	0.301	0.226
3	3.4	0.477	0.534
4	5.7	0.602	0.753
5	8.4	0.699	0.922

FIGURE 17.10
(*a*) Plot of untransformed data with the power equation that fits these data. (*b*) Plot of transformed data used to determine the coefficients of the power equation.