Thus, the intercept, $\log \alpha_2$, equals -0.300, and therefore, by taking the antilogarithm, $\alpha_2 = 10^{-0.3} = 0.5$. The slope is $\beta_2 = 1.75$. Consequently, the power equation is

$$y = 0.5x^{1.75}$$

This curve, as plotted in Fig. 17.10*a*, indicates a good fit.

17.1.6 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 (try Prob. 17.4 at the end of the chapter). 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.

17.2 POLYNOMIAL REGRESSION

In Sec. 17.1, a procedure was developed to derive the equation of a straight line using the least-squares criterion. Some engineering data, although exhibiting a marked pattern such as seen in Fig. 17.8, is poorly represented by a straight line. For these cases, a curve would be better suited to fit these data. As discussed in the previous section, one method to accomplish this objective is to use transformations. Another alternative is to fit polynomials to the data using *polynomial regression.*

The least-squares procedure can be readily extended to fit the data to a higher-order polynomial. For example, suppose that we fit a second-order polynomial or quadratic:

$$y = a_0 + a_1x + a_2x^2 + e$$

For this case the sum of the squares of the residuals is [compare with Eq. (17.3)]

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1x_i - a_2x_i^2)^2 \tag{17.18}$$

Following the procedure of the previous section, we take the derivative of Eq. (17.18) with respect to each of the unknown coefficients of the polynomial, as in

$$\frac{\partial S_r}{\partial a_0} = -2\sum (y_i - a_0 - a_1x_i - a_2x_i^2)$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum x_i(y_i - a_0 - a_1x_i - a_2x_i^2)$$

$$\frac{\partial S_r}{\partial a_2} = -2\sum x_i^2(y_i - a_0 - a_1x_i - a_2x_i^2)$$

These equations can be set equal to zero and rearranged to develop the following set of normal equations:

$$\begin{aligned}(n)a_0 + \left(\sum x_i\right)a_1 + \left(\sum x_i^2\right)a_2 &= \sum y_i\\ \left(\sum x_i\right)a_0 + \left(\sum x_i^2\right)a_1 + \left(\sum x_i^3\right)a_2 &= \sum x_iy_i\\ \left(\sum x_i^2\right)a_0 + \left(\sum x_i^3\right)a_1 + \left(\sum x_i^4\right)a_2 &= \sum x_i^2y_i\end{aligned} \tag{17.19}$$

where all summations are from $i = 1$ through n. Note that the above three equations are linear and have three unknowns: a_0, a_1, and a_2. The coefficients of the unknowns can be calculated directly from the observed data.

For this case, we see that the problem of determining a least-squares second-order polynomial is equivalent to solving a system of three simultaneous linear equations. Techniques to solve such equations were discussed in Part Three.

The two-dimensional case can be easily extended to an mth-order polynomial as

$$y = a_0 + a_1x + a_2x^2 + \cdots + a_mx^m + e$$

The foregoing analysis can be easily extended to this more general case. Thus, we can recognize that determining the coefficients of an mth-order polynomial is equivalent to solving a system of $m + 1$ simultaneous linear equations. For this case, the standard error is formulated as

$$s_{y/x} = \sqrt{\frac{S_r}{n - (m + 1)}} \tag{17.20}$$

This quantity is divided by $n - (m + 1)$ because $(m + 1)$ data-derived coefficients—a_0, a_1, . . . , a_m—were used to compute S_r; thus, we have lost $m + 1$ degrees of freedom. In addition to the standard error, a coefficient of determination can also be computed for polynomial regression with Eq. (17.10).

EXAMPLE 17.5 Polynomial Regression

Problem Statement. Fit a second-order polynomial to the data in the first two columns of Table 17.4.

Solution. From the given data,

$$\begin{array}{lll} m = 2 & \sum x_i = 15 & \sum x_i^4 = 979\\ n = 6 & \sum y_i = 152.6 & \sum x_iy_i = 585.6\\ \bar{x} = 2.5 & \sum x_i^2 = 55 & \sum x_i^2y_i = 2488.8\\ \bar{y} = 25.433 & \sum x_i^3 = 225 & \end{array}$$

TABLE 17.4 Computations for an error analysis of the quadratic least-squares fit.

x_i	y_i	$(y_i - \bar{y})^2$	$(y_i - a_0 - a_1x_i - a_2x_i^2)^2$
0	2.1	544.44	0.14332
1	7.7	314.47	1.00286
2	13.6	140.03	1.08158
3	27.2	3.12	0.80491
4	40.9	239.22	0.61951
5	61.1	1272.11	0.09439
Σ	152.6	2513.39	3.74657

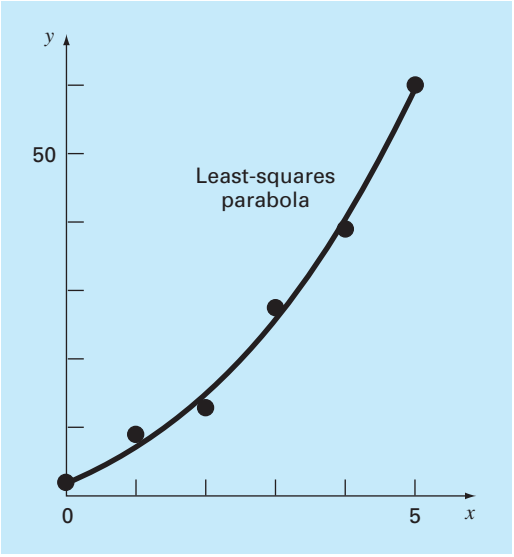

FIGURE 17.11
Fit of a second-order polynomial.

Therefore, the simultaneous linear equations are

$$\begin{bmatrix} 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} 152.6 \\ 585.6 \\ 2488.8 \end{Bmatrix}$$

Solving these equations through a technique such as Gauss elimination gives $a_0 = 2.47857$, $a_1 = 2.35929$, and $a_2 = 1.86071$. Therefore, the least-squares quadratic equation for this case is

$$y = 2.47857 + 2.35929x + 1.86071x^2$$

The standard error of the estimate based on the regression polynomial is [Eq. (17.20)]

$$s_{y/x} = \sqrt{\frac{3.74657}{6 - 3}} = 1.12$$

The coefficient of determination is

$$r^2 = \frac{2513.39 - 3.74657}{2513.39} = 0.99851$$

and the correlation coefficient is $r = 0.99925$.

These results indicate that 99.851 percent of the original uncertainty has been explained by the model. This result supports the conclusion that the quadratic equation represents an excellent fit, as is also evident from Fig. 17.11.

4th Polynomial Regression for Curve Fitting

A polynomial of the fourth order can be written as:

$$f(x) = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$$

Curve fitting of the data points with this polynomial is done by polynomial regression. The values of the five coefficients a_0, a_1, a_2, a_3, and a_4 are obtained by solving a system of five linear equations. The five equations can be written by extending Eqs.

$$na_0 + \left(\sum_{i=1}^{n} x_i\right)a_1 + \left(\sum_{i=1}^{n} x_i^2\right)a_2 + \left(\sum_{i=1}^{n} x_i^3\right)a_3 + \left(\sum_{i=1}^{n} x_i^4\right)a_4 = \sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} x_i\right)a_0 + \left(\sum_{i=1}^{n} x_i^2\right)a_1 + \left(\sum_{i=1}^{n} x_i^3\right)a_2 + \left(\sum_{i=1}^{n} x_i^4\right)a_3 + \left(\sum_{i=1}^{n} x_i^5\right)a_4 = \sum_{i=1}^{n} x_iy_i$$

$$\left(\sum_{i=1}^{n} x_i^2\right)a_0 + \left(\sum_{i=1}^{n} x_i^3\right)a_1 + \left(\sum_{i=1}^{n} x_i^4\right)a_2 + \left(\sum_{i=1}^{n} x_i^5\right)a_3 + \left(\sum_{i=1}^{n} x_i^6\right)a_4 = \sum_{i=1}^{n} x_i^2y_i$$

$$\left(\sum_{i=1}^{n} x_i^3\right)a_0 + \left(\sum_{i=1}^{n} x_i^4\right)a_1 + \left(\sum_{i=1}^{n} x_i^5\right)a_2 + \left(\sum_{i=1}^{n} x_i^6\right)a_3 + \left(\sum_{i=1}^{n} x_i^7\right)a_4 = \sum_{i=1}^{n} x_i^3y_i$$

$$\left(\sum_{i=1}^{n} x_i^4\right)a_0 + \left(\sum_{i=1}^{n} x_i^5\right)a_1 + \left(\sum_{i=1}^{n} x_i^6\right)a_2 + \left(\sum_{i=1}^{n} x_i^7\right)a_3 + \left(\sum_{i=1}^{n} x_i^8\right)a_4 = \sum_{i=1}^{n} x_i^4y_i$$

PROBLEMS

17.3 Use least-squares regression to fit a straight line to

x	0	2	4	6	9	11	12	15	17	19
y	5	6	7	6	9	8	7	10	12	12

Along with the slope and intercept, compute the standard error of the estimate and the correlation coefficient. Plot the data and the regression line. Then repeat the problem, but regress x versus y—that is, switch the variables. Interpret your results.

17.4 Use least-squares regression to fit a straight line to

x	6	7	11	15	17	21	23	29	29	37	39
y	29	21	29	14	21	15	7	7	13	0	3

Along with the slope and the intercept, compute the standard error of the estimate and the correlation coefficient. Plot the data and the regression line. If someone made an additional measurement of $x = 10$, $y = 10$, would you suspect, based on a visual assessment and the standard error, that the measurement was valid or faulty? Justify your conclusion.

17.5 Using the same approach as was employed to derive Eqs. (17.15) and (17.16), derive the least-squares fit of the following model:

$$y = a_1 x + e$$

That is, determine the slope that results in the least-squares fit for a straight line with a zero intercept. Fit the following data with this model and display the result graphically:

x	2	4	6	7	10	11	14	17	20
y	1	2	5	2	8	7	6	9	12

17.6 Use least-squares regression to fit a straight line to

x	1	2	3	4	5	6	7	8	9
y	1	1.5	2	3	4	5	8	10	13

(a) Along with the slope and intercept, compute the standard error of the estimate and the correlation coefficient. Plot the data and the straight line. Assess the fit.

(b) Recompute **(a)**, but use polynomial regression to fit a parabola to the data. Compare the results with those of **(a)**.

17.7 Fit the following data with **(a)** a saturation-growth-rate model, **(b)** a power equation, and **(c)** a parabola. In each case, plot the data and the equation.

x	0.75	2	3	4	6	8	8.5
y	1.2	1.95	2	2.4	2.4	2.7	2.6

17.8 Fit the following data with the power model ($y = ax^b$). Use the resulting power equation to predict y at $x = 9$:

x	2.5	3.5	5	6	7.5	10	12.5	15	17.5	20
y	13	11	8.5	8.2	7	6.2	5.2	4.8	4.6	4.3

17.9 Fit an exponential model to

x	0.4	0.8	1.2	1.6	2	2.3
y	800	975	1500	1950	2900	3600

Plot the data and the equation on both standard and semi-logarithmic graph paper.

17.10 Rather than using the base-e exponential model (Eq. 17.22), a common alternative is to use a base-10 model,

$$y = \alpha_5 10^{\beta_5 x}$$

When used for curve fitting, this equation yields identical results to the base-e version, but the value of the exponent parameter (β_5) will differ from that estimated with Eq. 17.22 (β_1). Use the base-10 version to solve Prob. 17.9. In addition, develop a formulation to relate β_1 to β_5.

17.11 Beyond the examples in Fig. 17.10, there are other models that can be linearized using transformations. For example,

$$y = \alpha_4 x e^{\beta_4 x}$$

Linearize this model and use it to estimate α_4 and β_4 based on the following data. Develop a plot of your fit along with the data.

x	0.1	0.2	0.4	0.6	0.9	1.3	1.5	1.7	1.8
y	0.75	1.25	1.45	1.25	0.85	0.55	0.35	0.28	0.18

17.12 An investigator has reported the data tabulated below for an experiment to determine the growth rate of bacteria k (per d), as a function of oxygen concentration c (mg/L). It is known that such data can be modeled by the following equation:

$$k = \frac{k_{max} c^2}{c_s + c^2}$$

where c_s and k_{max} are parameters. Use a transformation to linearize this equation. Then use linear regression to estimate c_s and k_{max} and predict the growth rate at $c = 2$ mg/L.

c	0.5	0.8	1.5	2.5	4
k	1.1	2.4	5.3	7.6	8.9

17.13 An investigator has reported the data tabulated below. It is known that such data can be modeled by the following equation

$$x = e^{(y-b)/a}$$

where a and b are parameters. Use a transformation to linearize this equation and then employ linear regression to determine a and b. Based on your analysis predict y at $x = 2.6$.

x	1	2	3	4	5
y	0.5	2	2.9	3.5	4

17.14 It is known that the data tabulated below can be modeled by the following equation

$$y = \left(\frac{a + \sqrt{x}}{b\sqrt{x}}\right)^2$$

Use a transformation to linearize this equation and then employ linear regression to determine the parameters a and b. Based on your analysis predict y at $x = 1.6$.

x	0.5	1	2	3	4
y	10.4	5.8	3.3	2.4	2

17.15 The following data are provided

x	1	2	3	4	5
y	2.2	2.8	3.6	4.5	5.5

You want to use least-squares regression to fit these data with the following model,

$$y = a + bx + \frac{c}{x}$$

Determine the coefficients by setting up and solving Eq. (17.25).

17.16 Given these data

x	5	10	15	20	25	30	35	40	45	50
y	17	24	31	33	37	37	40	40	42	41

use least-squares regression to fit **(a)** a straight line, **(b)** a power equation, **(c)** a saturation-growth-rate equation, and **(d)** a parabola. Plot the data along with all the curves. Is any one of the curves superior? If so, justify.

17.17 Fit a cubic equation to the following data:

x	3	4	5	7	8	9	11	12
y	1.6	3.6	4.4	3.4	2.2	2.8	3.8	4.6

Along with the coefficients, determine r^2 and $s_{y/x}$.

17.18 Use multiple linear regression to fit

x_1	0	1	1	2	2	3	3	4	4
x_2	0	1	2	1	2	1	2	1	2
y	15.1	17.9	12.7	25.6	20.5	35.1	29.7	45.4	40.2

Compute the coefficients, the standard error of the estimate, and the correlation coefficient.

17.19 Use multiple linear regression to fit

x_1	0	0	1	2	0	1	2	2	1
x_2	0	2	2	4	4	6	6	2	1
y	14	21	11	12	23	23	14	6	11

Compute the coefficients, the standard error of the estimate, and the correlation coefficient.

17.20 Use nonlinear regression to fit a parabola to the following data:

x	0.2	0.5	0.8	1.2	1.7	2	2.3
y	500	700	1000	1200	2200	2650	3750

17.21 Use nonlinear regression to fit a saturation-growth-rate equation to the data in Prob. 17.16.

17.22 Recompute the regression fits from Probs. **(a)** 17.3 and **(b)** 17.17, using the matrix approach. Estimate the standard errors and develop 90% confidence intervals for the coefficients.

17.23 Develop, debug, and test a program in either a high-level language or macro language of your choice to implement linear regression. Among other things: **(a)** include statements to document the code, and **(b)** determine the standard error and the coefficient of determination.

17.24 A material is tested for cyclic fatigue failure whereby a stress, in MPa, is applied to the material and the number of cycles needed to cause failure is measured. The results are in the table below. When a log-log plot of stress versus cycles is generated, the

data trend shows a linear relationship. Use least-squares regression to determine a best-fit equation for these data.

N, cycles	1	10	100	1000	10,000	100,000	1,000,000
Stress, MPa	1100	1000	925	800	625	550	420

17.25 The following data show the relationship between the viscosity of SAE 70 oil and temperature. After taking the log of the data, use linear regression to find the equation of the line that best fits the data and the r^2 value.

Temperature, °C	26.67	93.33	148.89	315.56
Viscosity, μ, N · s/m^2	1.35	0.085	0.012	0.00075

17.26 The data below represents the bacterial growth in a liquid culture over a number of days.

Day	0	4	8	12	16	20
Amount × 10^6	67	84	98	125	149	185

Find a best-fit equation to the data trend. Try several possibilities—linear, parabolic, and exponential. Use the software package of your choice to find the best equation to predict the amount of bacteria after 40 days.

17.27 The concentration of *E. coli* bacteria in a swimming area is monitored after a storm:

t (hr)	4	8	12	16	20	24
c (CFU/100 mL)	1600	1320	1000	890	650	560

The time is measured in hours following the end of the storm and the unit CFU is a "colony forming unit." Use these data to estimate **(a)** the concentration at the end of the storm ($t = 0$) and **(b)** the time at which the concentration will reach 200 CFU/100 mL. Note that your choice of model should be consistent with the fact that negative concentrations are impossible and that the bacteria concentration always decreases with time.

17.28 An object is suspended in a wind tunnel and the force measured for various levels of wind velocity. The results are tabulated below.

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

Use least-squares regression to fit these data with **(a)** a straight line, **(b)** a power equation based on log transformations, and **(c)** a power model based on nonlinear regression. Display the results graphically.

17.29 Fit a power model to the data from Prob. 17.28, but use natural logarithms to perform the transformations.

17.30 Derive the least-squares fit of the following model:

$$y = a_1 x + a_2 x^2 + e$$

That is, determine the coefficients that results in the least-squares fit for a second-order polynomial with a zero intercept. Test the approach by using it to fit the data from Prob. 17.28.

17.31 In Prob. 17.11 we used transformations to linearize and fit the following model:

$$y = \alpha_4 x e^{\beta_4 x}$$

Use nonlinear regression to estimate α_4 and β_4 based on the following data. Develop a plot of your fit along with the data.

x	0.1	0.2	0.4	0.6	0.9	1.3	1.5	1.7	1.8
y	0.75	1.25	1.45	1.25	0.85	0.55	0.35	0.28	0.18