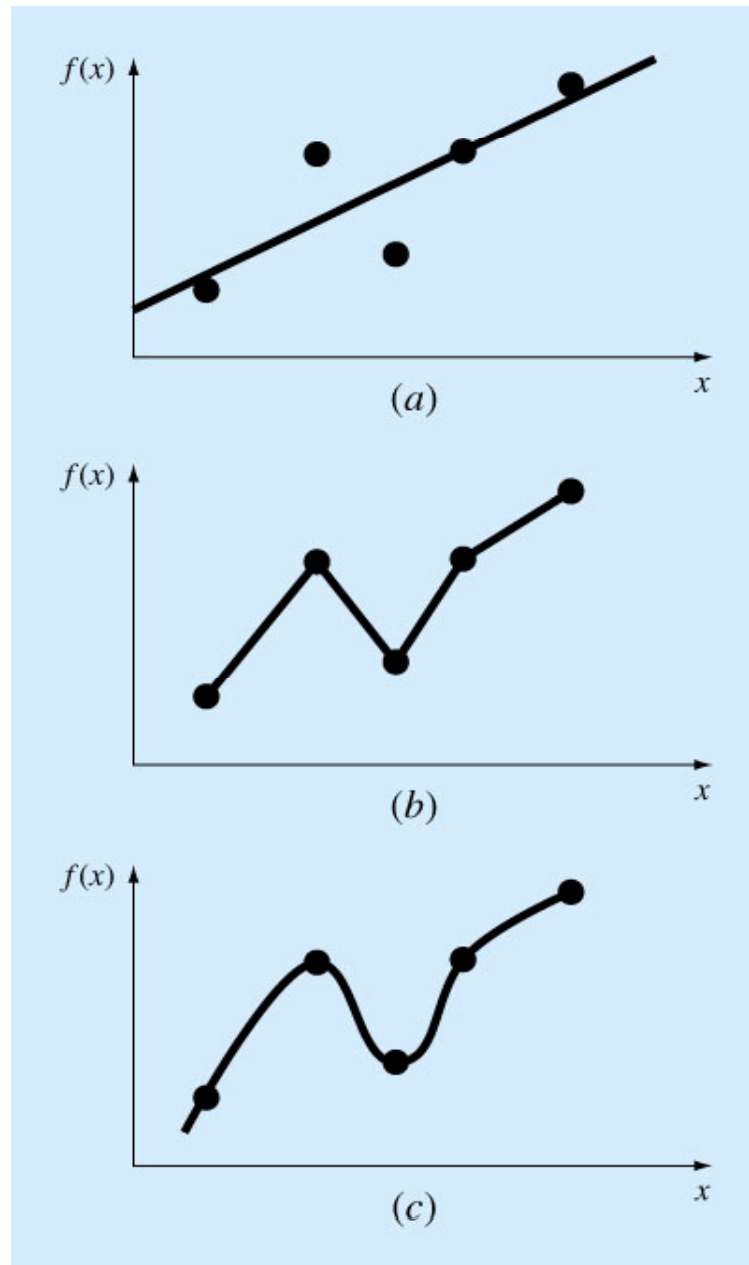


# CURVE FITTING

- Describes techniques to fit curves (*curve fitting*) to discrete data to obtain intermediate estimates.
- There are two general approaches to curve fitting:
  - *Data exhibit a significant degree of scatter.* The strategy is to derive a single curve that represents the general trend of the data.
  - *Data is very precise.* The strategy is to pass a curve or a series of curves through each of the points.
- In engineering two types of applications are encountered:
  - Trend analysis. Predicting values of dependent variable, may include extrapolation beyond data points or interpolation between data points.
  - Hypothesis testing. Comparing existing mathematical model with measured data.

Figure PT5.1



# Mathematical Background

## Simple Statistics/

- In course of engineering study, if several measurements are made of a particular quantity, 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
  - The location of the center of the distribution of the data,
  - The degree of spread of the data.

- *Arithmetic mean*. The sum of the individual data points ( $y_i$ ) divided by the number of points ( $n$ ).

$$\bar{y} = \frac{\sum_{i=1, \dots, n} y_i}{n}$$

- *Standard deviation*. The most common measure of a spread for a sample.

$$S_y = \sqrt{\frac{S_t}{n-1}}$$
$$S_t = \sum (y_i - \bar{y})^2$$

or

$$S_y^2 = \frac{\sum y_i^2 - (\sum y_i)^2 / n}{n-1}$$

- *Variance*. Representation of spread by the square of the standard deviation.

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

Degrees of freedom

- *Coefficient of variation*. Has the utility to quantify the spread of data.

$$c.v. = \frac{s_y}{\bar{y}} 100\%$$

Figure PT5.2

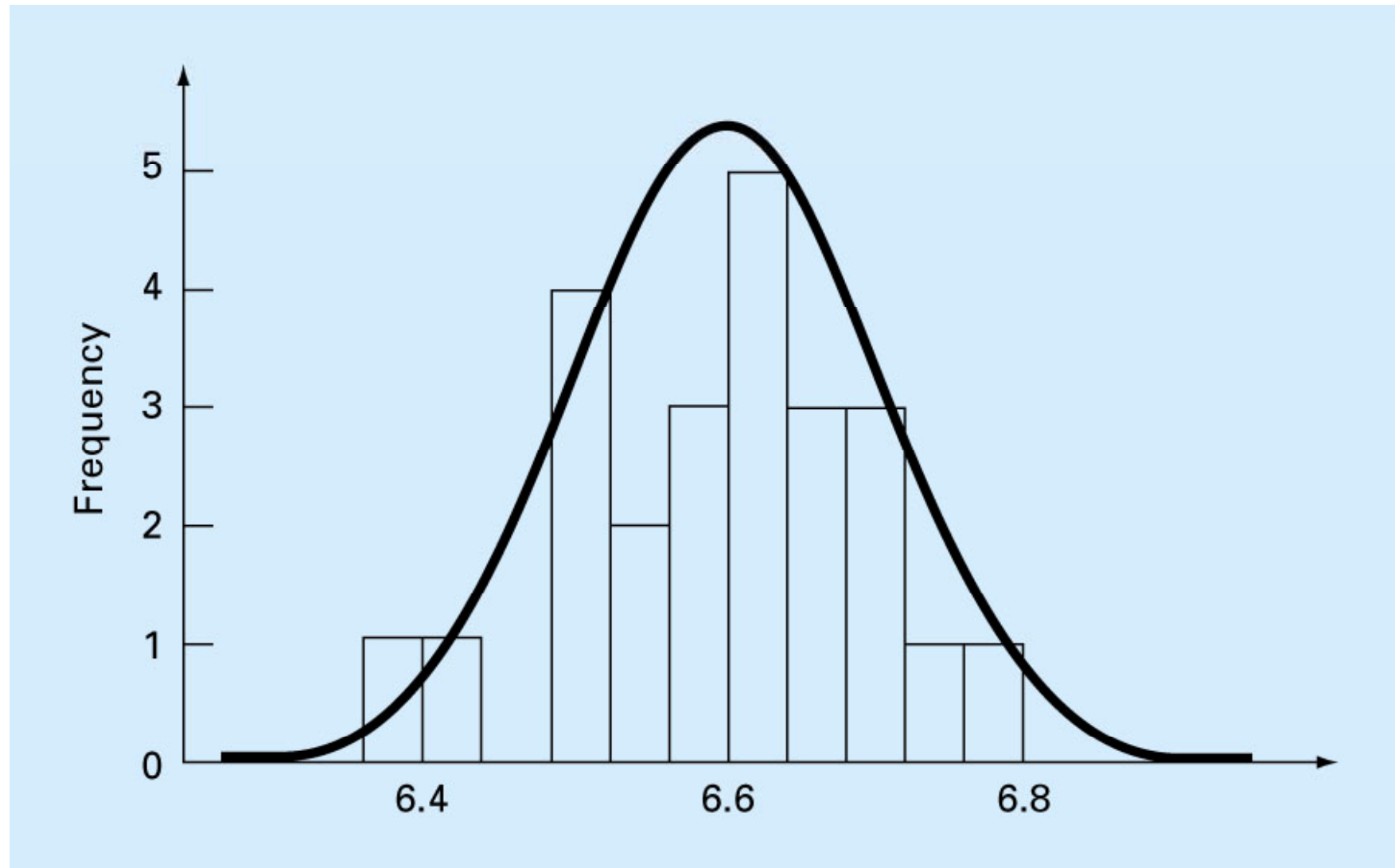
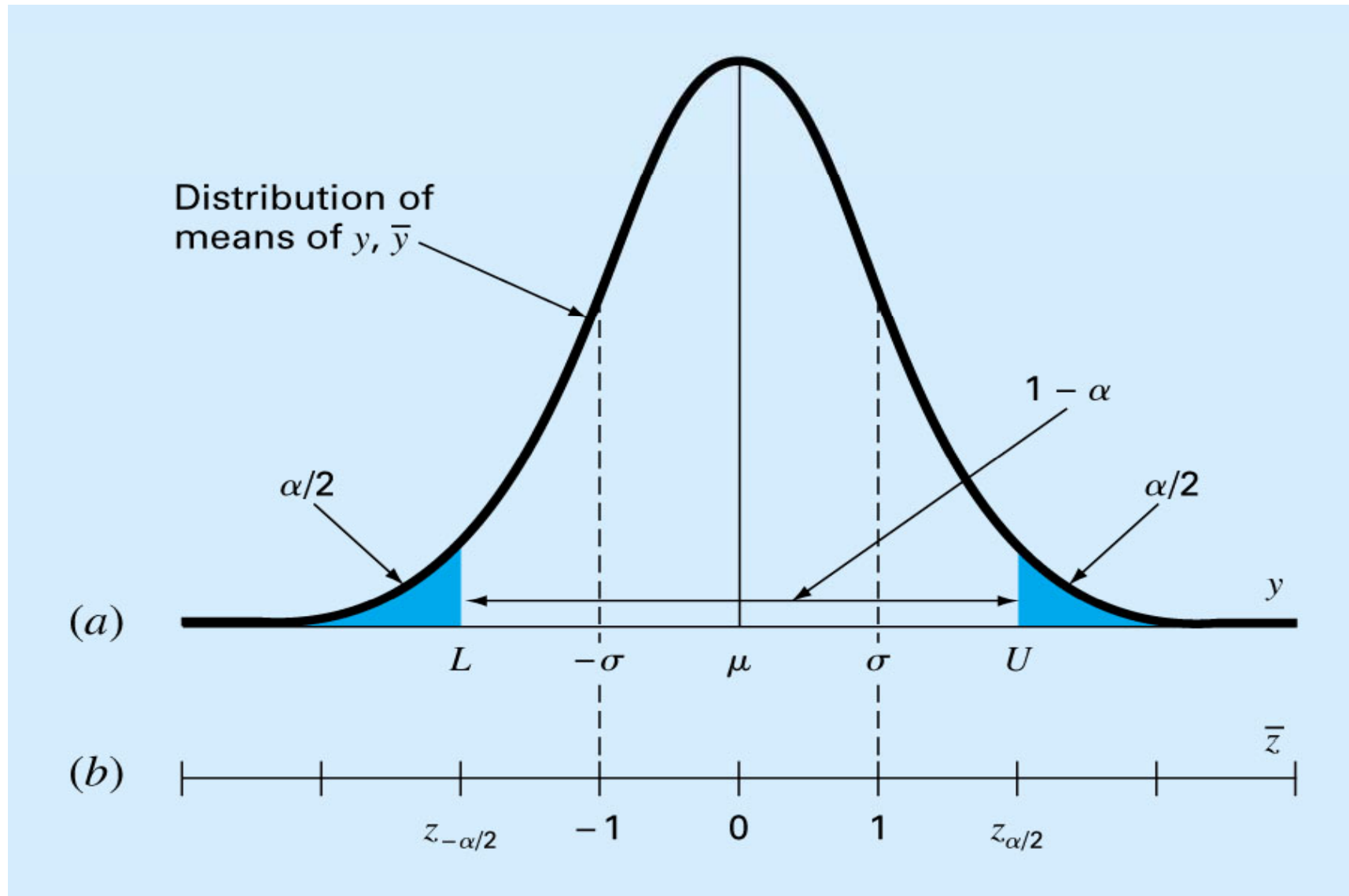


Figure PT5.3



# Least Squares Regression

## Chapter 17

### Linear Regression

- Fitting a straight line to a set of paired observations:  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ .

$$y = a_0 + a_1x + e$$

$a_1$ - slope

$a_0$ - intercept

$e$ - error, or residual, between the model and the observations



## Criteria for a “Best” Fit/

- Minimize the sum of the residual errors for all available data:

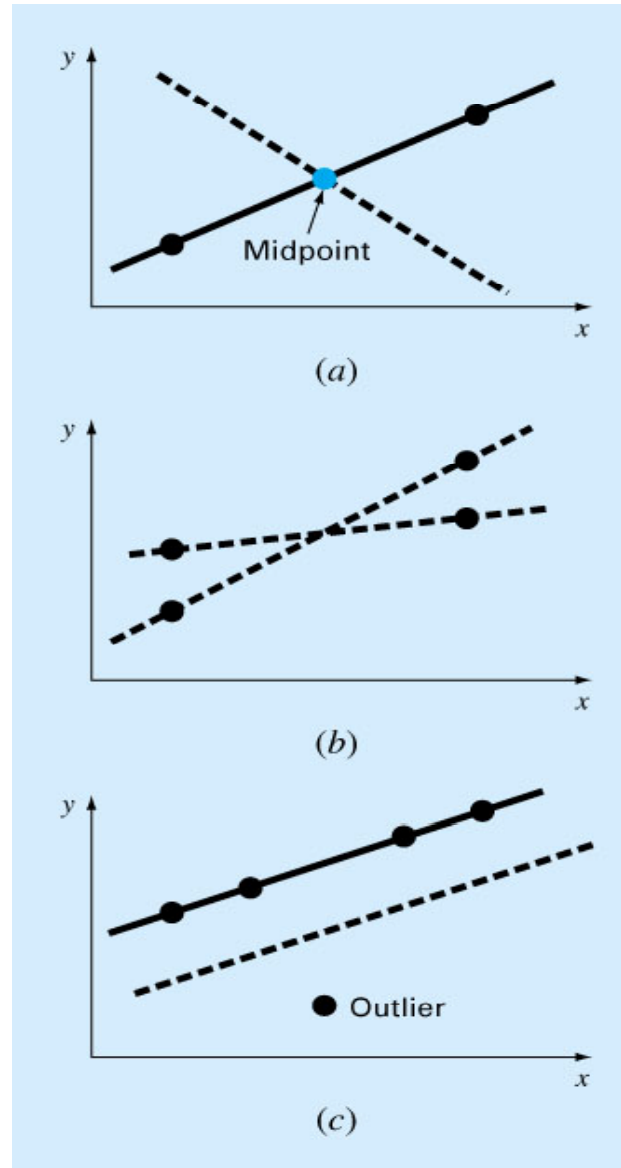
$$\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - a_o - a_1 x_i)$$

$n$  = total number of points

- However, this is an inadequate criterion, so is the sum of the absolute values

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

Figure 17.2



- Best strategy is to minimize the sum of the squares of the residuals between the measured  $y$  and the  $y$  calculated with the linear model:

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

- Yields a unique line for a given set of data.

## Least-Squares Fit of a Straight Line/

$$\frac{\partial S_r}{\partial a_o} = -2 \sum (y_i - a_o - a_1 x_i) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum [(y_i - a_o - a_1 x_i) x_i] = 0$$

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

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

$$\left. \begin{aligned} \sum a_o &= n a_o \\ n a_o + \left( \sum x_i \right) a_1 &= \sum y_i \end{aligned} \right\} \text{Normal equations, can be solved simultaneously}$$

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

$$a_o = \bar{y} - a_1 \bar{x}$$

Mean values

Chapter 17

Figure 17.3

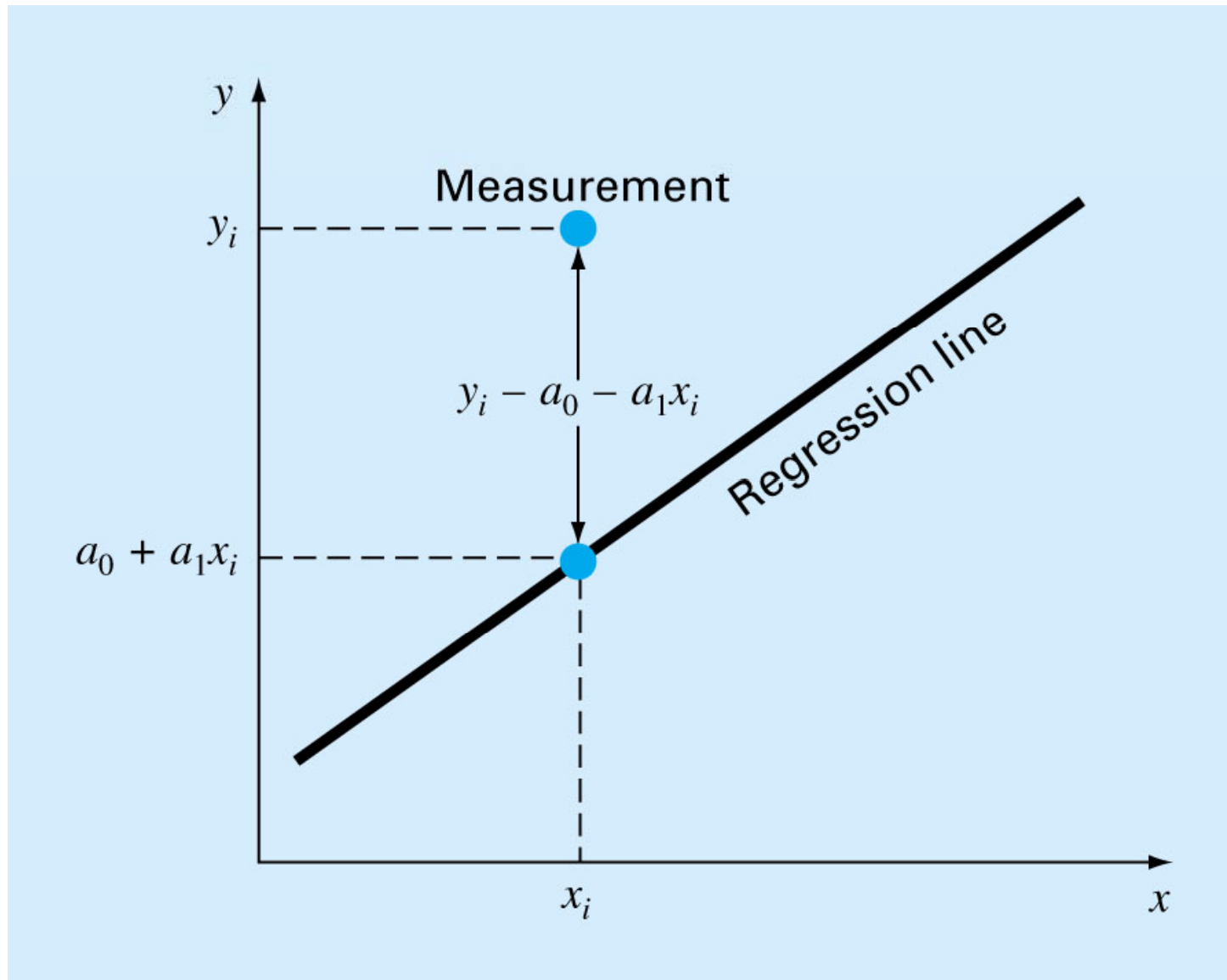


Figure 17.4

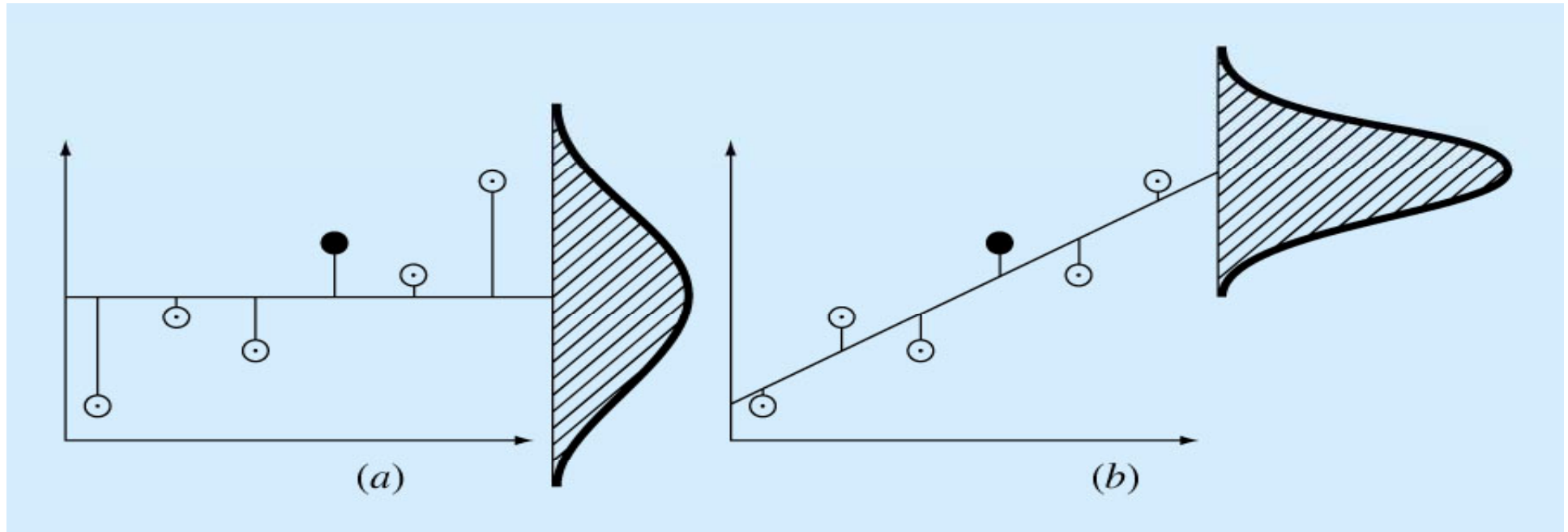
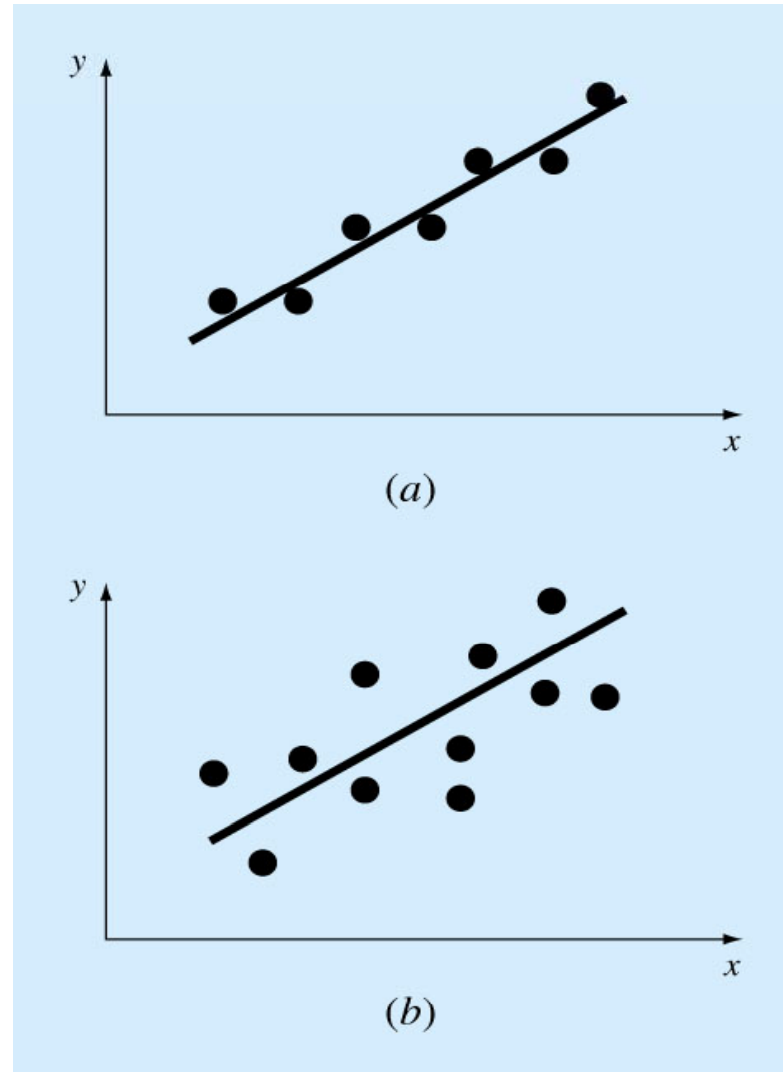


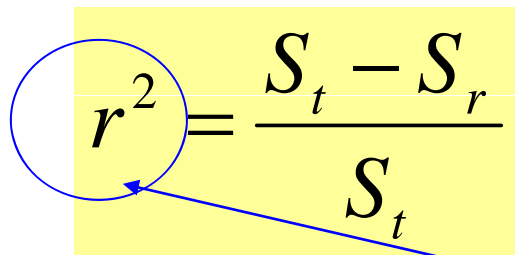
Figure 17.5



## “Goodness” of our fit/

If

- Total sum of the squares around the mean for the dependent variable,  $y$ , is  $S_t$
- Sum of the squares of residuals around the regression line is  $S_r$
- $S_t - S_r$  quantifies the improvement or error reduction due to describing data in terms of a straight line rather than as an average value.


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

$r^2$ -coefficient of determination

Sqrt( $r^2$ ) – correlation coefficient



- For a perfect fit  
 $S_r=0$  and  $r=r^2=1$ , signifying that the line explains 100 percent of the variability of the data.
- For  $r=r^2=0$ ,  $S_r=S_t$ , the fit represents no improvement.

## 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:

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

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

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

Using Eqs. (17.6) and (17.7),

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

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

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

| $x_i$    | $y_i$ | $(y_i - \bar{y})^2$ | $(y_i - a_0 - a_1 x_i)^2$ |
|----------|-------|---------------------|---------------------------|
| 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                    |
| $\Sigma$ | 24.0  | 22.7143             | 2.9911                    |

Therefore, the least-squares fit is

$$y = 0.07142857 + 0.8392857x$$

### 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 data in Example 17.1.

**Solution.** The summations are performed and presented in Table 17.1. The standard deviation is [Eq. (PT5.2)]

$$s_y = \sqrt{\frac{22.7143}{7-1}} = 1.9457$$

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

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

Thus, because  $s_{y/x} < s_y$ , the linear regression model has merit. The extent of the improvement is quantified by [Eq. (17.10)]

$$r^2 = \frac{22.7143 - 2.9911}{22.7143} = 0.868$$

or

$$r = \sqrt{0.868} = 0.932$$

These results indicate that 86.8 percent of the original uncertainty has been explained by the linear model.

```
SUB Regress(x, y, n, a1, a0, syx, r2)
```

```
    sumx = 0: sumxy = 0: st = 0
```

```
    sumy = 0: sumx2 = 0: sr = 0
```

```
    DOFOR i = 1, n
```

```
        sumx = sumx + xi
```

```
        sumy = sumy + yi
```

```
        sumxy = sumxy + xi*yi
```

```
        sumx2 = sumx2 + xi*xi
```

```
    END DO
```

```
    xm = sumx/n
```

```
    ym = sumy/n
```

```
    a1 = (n*sumxy - sumx*sumy)/(n*sumx2 - sumx*sumx)
```

```
    a0 = ym - a1*xm
```

```
    DOFOR i = 1, n
```

```
        st = st + (yi - ym)2
```

```
        sr = sr + (yi - a1*xi - a0)2
```

```
    END DO
```

```
    syx = (sr/(n - 2))0.5
```

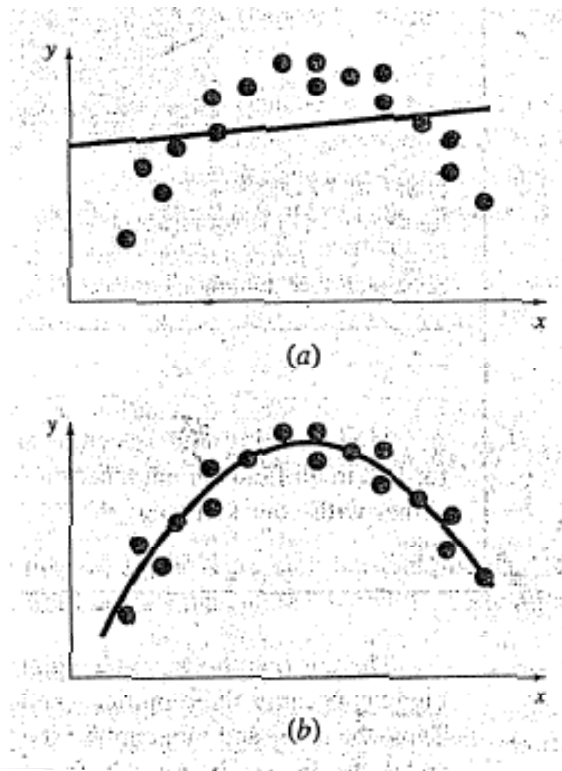
```
    r2 = (st - sr)/st
```

```
END Regress
```

# Linearization of nonlinear relationships

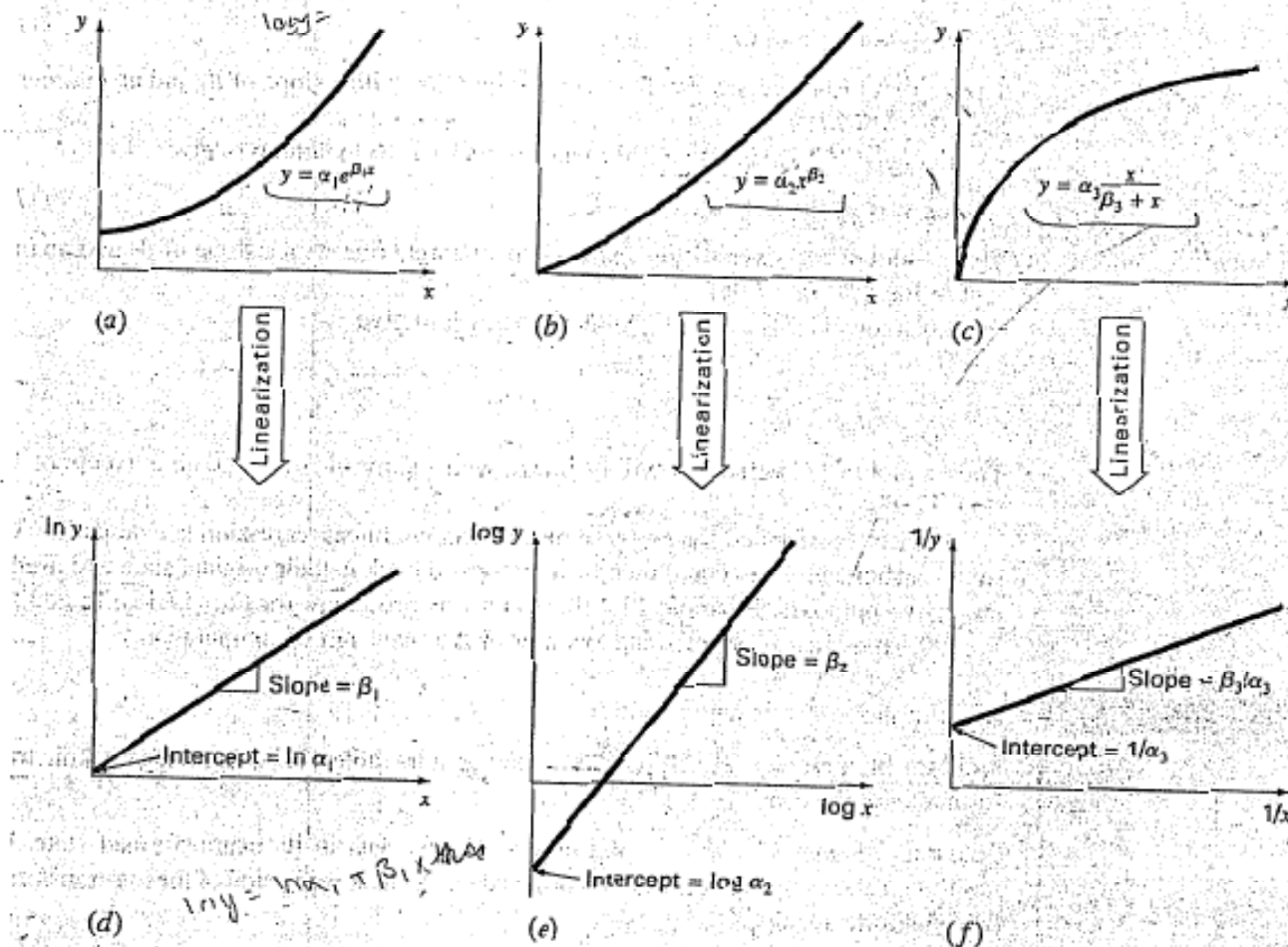
**FIGURE 17.8**

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



**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.



# Polynomial Regression

- Some engineering data is poorly represented by a straight line. For these cases a curve is better suited to fit the data. The least squares method can readily be extended to fit the data to higher order polynomials ([Sec. 17.2](#)).

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)]

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

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:

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

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.



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

Error estandard

### Polynomial Regression

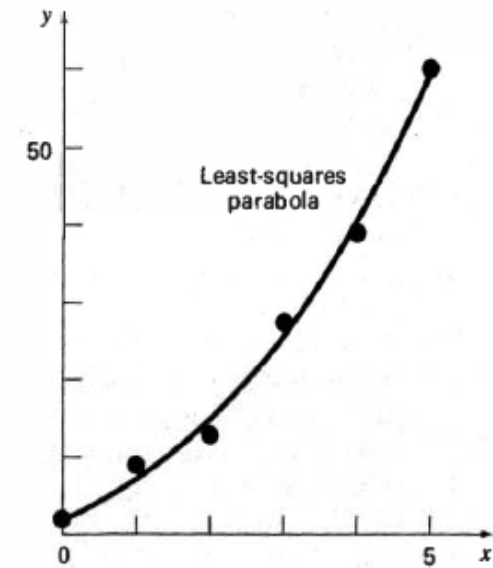
**Problem Statement.** Fit a second-order polynomial to the data in the first two columns of Table 17.4.

**Solution.** From the given data,

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

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

| $x_i$    | $y_i$ | $(y_i - \bar{y})^2$ | $(y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$ |
|----------|-------|---------------------|---------------------------------------|
| 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                               |
| $\Sigma$ | 152.6 | 2513.39             | 3.74657                               |



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.

#### FIGURE 17.12

Algorithm for implementation of polynomial and multiple linear regression.

- Step 1:** Input order of polynomial to be fit,  $m$ .
- Step 2:** Input number of data points,  $n$ .
- Step 3:** If  $n < m + 1$ , print out an error message that regression is impossible and terminate the process. If  $n \geq m + 1$ , continue.
- Step 4:** Compute the elements of the normal equation in the form of an augmented matrix.
- Step 5:** Solve the augmented matrix for the coefficients  $a_0, a_1, a_2, \dots, a_m$ , using an elimination method.
- Step 6:** Print out the coefficients.

# Multiple linear regression

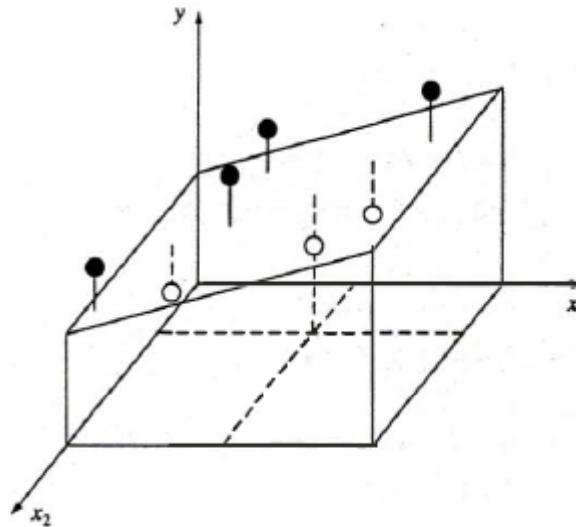
A useful extension of linear regression is the case where  $y$  is a linear function of two or more independent variables. For example,  $y$  might be a linear function of  $x_1$  and  $x_2$ , as in

$$y = a_0 + a_1x_1 + a_2x_2 + e$$

Such an equation is particularly useful when fitting experimental data where the variable being studied is often a function of two other variables. For this two-dimensional case, the regression “line” becomes a “plane” (Fig. 17.14).

**FIGURE 17.14**

Graphical depiction of multiple linear regression where  $y$  is a linear function of  $x_1$  and  $x_2$ .



As with the previous cases, the “best” values of the coefficients are determined by setting up the sum of the squares of the residuals,

$$S_r = \sum_{i=1}^n (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2 \quad (17.21)$$

and differentiating with respect to each of the unknown coefficients,

$$\begin{aligned} \frac{\partial S_r}{\partial a_0} &= -2 \sum (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i}) \\ \frac{\partial S_r}{\partial a_1} &= -2 \sum x_{1i} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i}) \\ \frac{\partial S_r}{\partial a_2} &= -2 \sum x_{2i} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i}) \end{aligned}$$

The coefficients yielding the minimum sum of the squares of the residuals are obtained by setting the partial derivatives equal to zero and expressing the result in matrix form as

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i} x_{2i} \\ \sum x_{2i} & \sum x_{1i} x_{2i} & \sum x_{2i}^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{1i} y_i \\ \sum x_{2i} y_i \end{bmatrix} \quad (17.22)$$

### EXAMPLE 17.6 Multiple Linear Regression

**Problem Statement.** The following data was calculated from the equation  $y = 5 + 4x_1 - 3x_2$ :

| $x_1$ | $x_2$ | $y$ |
|-------|-------|-----|
| 0     | 0     | 5   |
| 2     | 1     | 10  |
| 2.5   | 2     | 9   |
| 1     | 3     | 0   |
| 4     | 6     | 3   |
| 7     | 2     | 27  |

Use multiple linear regression to fit this data.

**Solution.** The summations required to develop Eq. (17.22) are computed in Table 17.5. The result is

$$\begin{bmatrix} 6 & 16.5 & 14 \\ 16.5 & 76.25 & 48 \\ 14 & 48 & 54 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} 54 \\ 243.5 \\ 100 \end{Bmatrix}$$

which can be solved using a method such as Gauss elimination for

$$a_0 = 5 \quad a_1 = 4 \quad a_2 = -3$$

which is consistent with the original equation from which the data was derived.

**TABLE 17.5** Computations required to develop the normal equations for Example 17.6

| $y$      | $x_1$ | $x_2$ | $x_1^2$ | $x_2^2$ | $x_1 x_2$ | $x_1 y$ |
|----------|-------|-------|---------|---------|-----------|---------|
| 5        | 0     | 0     | 0       | 0       | 0         | 0       |
| 10       | 2     | 1     | 4       | 1       | 2         | 20      |
| 9        | 2.5   | 2     | 6.25    | 4       | 5         | 22.5    |
| 0        | 1     | 3     | 1       | 9       | 3         | 0       |
| 3        | 4     | 6     | 16      | 36      | 24        | 12      |
| 27       | 7     | 2     | 49      | 4       | 14        | 189     |
| $\Sigma$ | 54    | 16.5  | 76.25   | 54      | 48        | 243.5   |

The foregoing two-dimensional case can be easily extended to  $m$  dimensions, as

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m + e$$

where the standard error is formulated as

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

and the coefficient of determination is computed as in Eq. (17.10). An algorithm to set up the normal equations is listed in Fig. 17.15.

Although there may be certain cases where a variable is linearly related to two or more other variables, multiple linear regression has additional utility in the derivation of power equations of the general form

$$y = a_0x_1^{a_1}x_2^{a_2}\cdots x_m^{a_m}$$

Such equations are extremely useful when fitting experimental data. To use multiple linear regression, the equation is transformed by taking its logarithm to yield

$$\log y = \log a_0 + a_1 \log x_1 + a_2 \log x_2 + \cdots + a_m \log x_m$$

# General Linear Least Squares

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + e$$

$z_0, z_1, \dots, z_m$  are  $m+1$  basis functions

$$\{Y\} = [Z]\{A\} + \{E\}$$

$[Z]$  – matrix of the calculated values of the basis functions  
at the measured values of the independent variable

$\{Y\}$  – observed values of the dependent variable

$\{A\}$  – unknown coefficients

$\{E\}$  – residuals

$$S_r = \sum_{i=1}^n \left( y_i - \sum_{j=0}^m a_j z_{ji} \right)^2$$

Minimized by taking its partial derivative w.r.t. each of the coefficients and setting the resulting equation equal to zero

Chapter 17