Curve Fitting

Part 5

- Describes techniques to fit curves (*curve fitting*) to discrete data to obtain intermediate estimates.
- There are two general approaches two 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.

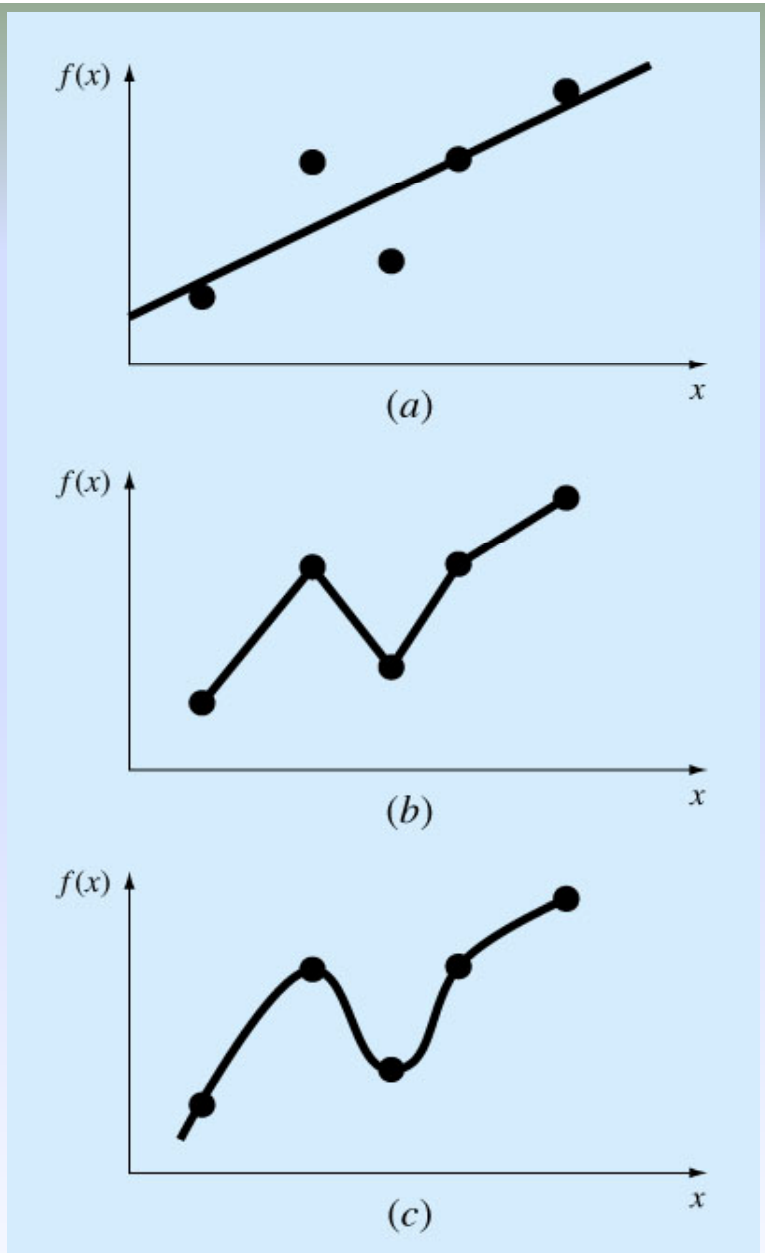


FIGURE PT5.1
Three attempts to fit a "best" curve through five data points. (*a*) Least-squares regression, (*b*) linear interpolation, and (*c*) curvilinear interpolation.

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.

Example

- Suppose that 24 measurements made of the coefficient of thermal expansion of a 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

Values change from 6.395 to 6.775. But we may use some statictical information about this data. These descriptive statistics are often selected to represent 1) the location of the center of the distribution of the data 2) the degree of spread of the data set.

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

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

$$i = 1, \ldots, n$$

- *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

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

Variance

- *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.

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

Example: continued

i	y_i	$(y_i-\bar{y})^2$
1	6.395	0.042025
2	6.435	0.027225
3	6.485	0.013225
4	6.495	0.011025
5	6.505	0.009025
6	6.515	0.007225
7	6.555	0.002025
8	6.555	0.002025
9	6.565	0.001225
10	6.575	0.000625
11	6.595	0.000025
12	6.605	0.000025
13	6.615	0.000225
14	6.625	0.000625
15	6.625	0.000625
16	6.635	0.001225
17	6.655	0.003025
18	6.655	0.003025
19	6.665	0.004225
20	6.685	0.007225
21	6.715	0.013225
22	6.715	0.013225
23	6.755	0.024025
24	6.775	0.030625

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

$$\sum (y_i - \bar{y})^2 = 0.217$$

$$s_y = \sqrt{\frac{0.217}{24-1}} = 0.097133$$

$$s_y^2 = 0.009435$$

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

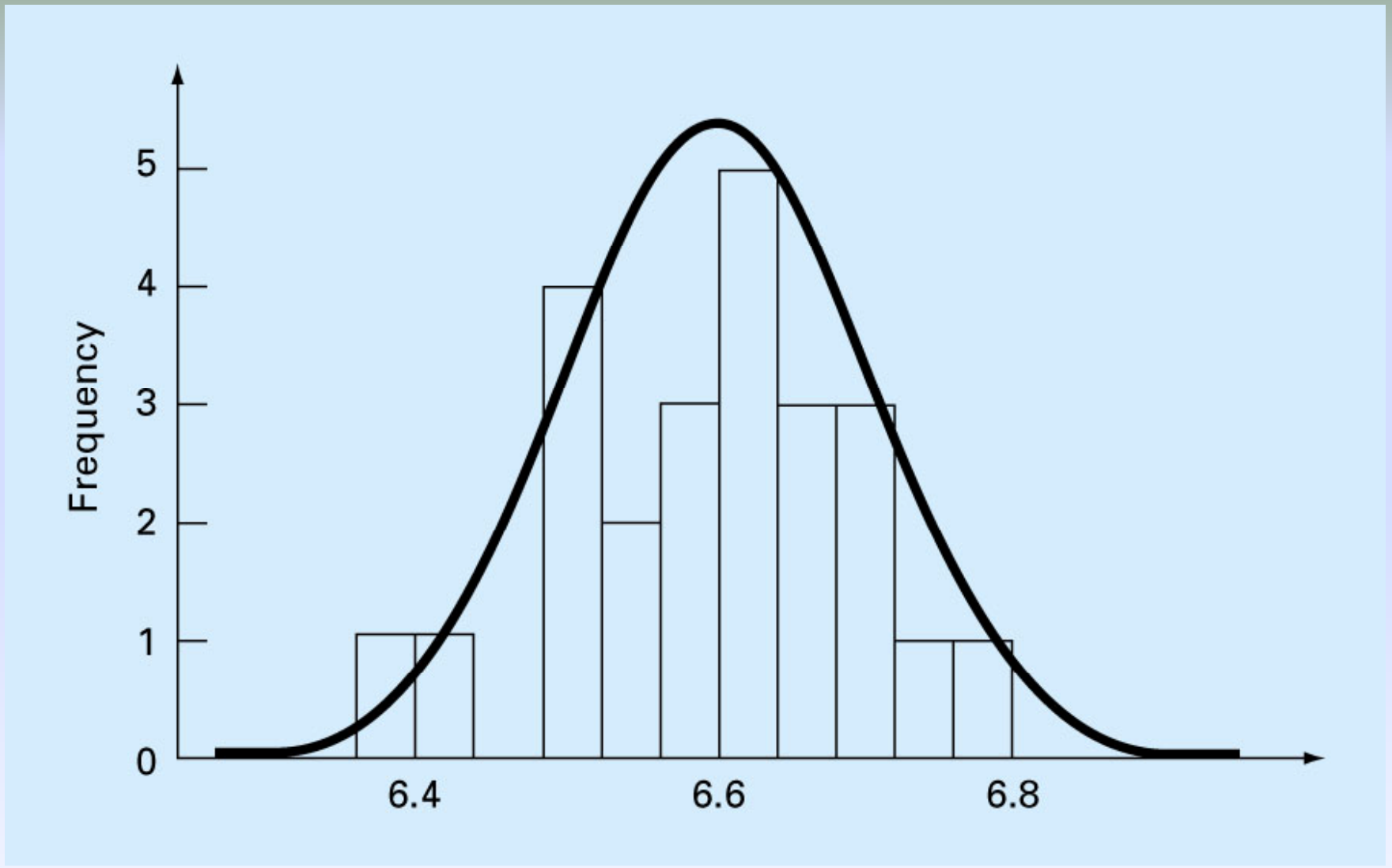


FIGURE PT5.2
A histogram used to depict the distribution of data. As the number of data points increases, the histogram could approach the smooth, bell-shaped curve called the normal distribution.

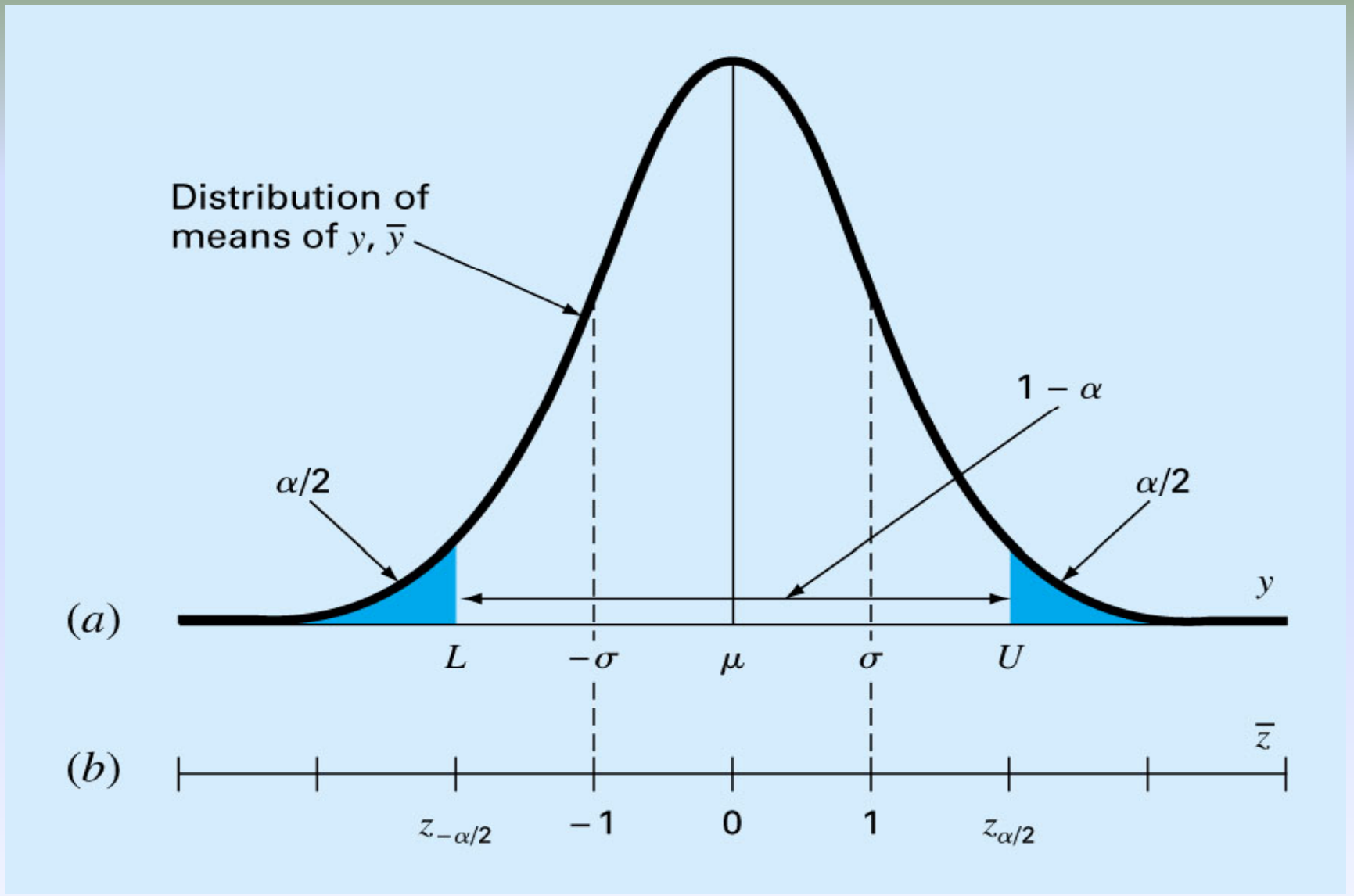


FIGURE PT5.3
A two-sided confidence interval. The abscissa scale in (*a*) is written in the natural units of the random variable *y*. The normalized version of the abscissa in (*b*) has the mean at the origin and scales the axis so that the standard deviation corresponds to a unit value.

Least Squares Regression

Chapter 17

Linear Regression

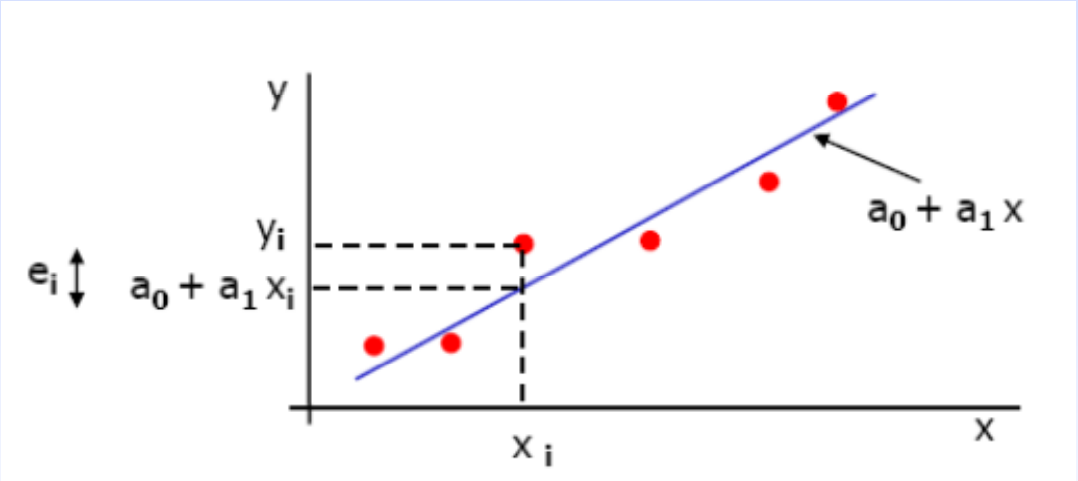
- Fitting a straight line to a set of paired observations: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

  $y = a_0 + a_1 x + e$

  a_1: slope

  a_0: intercept

  e: error



Criteria for a "best" Fit/

- *Strategy-1*: 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

- *Strategy-2*: minimize the sum of the absolute values

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

- *Strategy-3*: minimize the sum of the squares of the residuals between the *measured y* and the *y calculated* with the linear model:

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

- The line is chosen that minimizes the maximum distance that an individual point falls from the line.
- This is the best strategy but it is not suitable in the existance of an outlier, that is a single point with a large error.

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

We need to find unknowns a_0 and a_1 to determine the line.

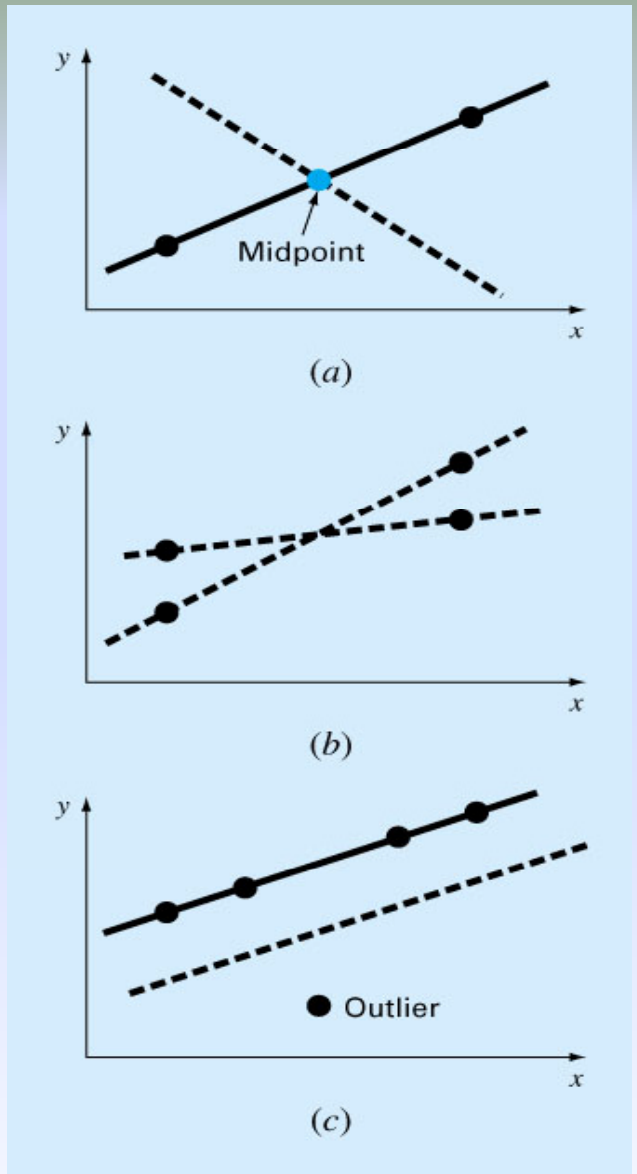


FIGURE 17.2
Examples of some criteria for "best fit" that are inadequate for regression: (*a*) minimizes the sum of the residuals, (*b*) minimizes the sum of the absolute values of the residuals, and (*c*) minimizes the maximum error of any individual point.

Least-Squares Fit of a Straight Line/

Best strategy: find the unknowns a_0 and a_1 by minimizing S_r :

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

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

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

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

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

$$\sum a_0 = n a_0$$

$$n a_0 + \left(\sum x_i\right) a_1 = \sum y_i$$

} Normal equations, can be solved simultaneously

Solution:

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

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

Mean values

Line:

$$y = a_0 + a_1 x$$

Example

Use least-squares regression to fit a straight line to

x	1	3	5	7	10	12	13	16	18	20
y	4	5	6	5	8	7	6	9	12	11

$n = 10$

$\sum x_i = 105$

$\sum y_i = 73$

$\bar{x} = 10.5$

$\bar{y} = 7.3$

$\sum x_i^2 = 1477$

$\sum x_i y_i = 906$

$$a_1 = \frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{10*906 - 105*73}{10*1477 - 105^2} = 0.3725$$

$$a_0 = 7.3 - 0.3725*10.5 = 3.3888$$

$$y = a_0 + a_1 x$$

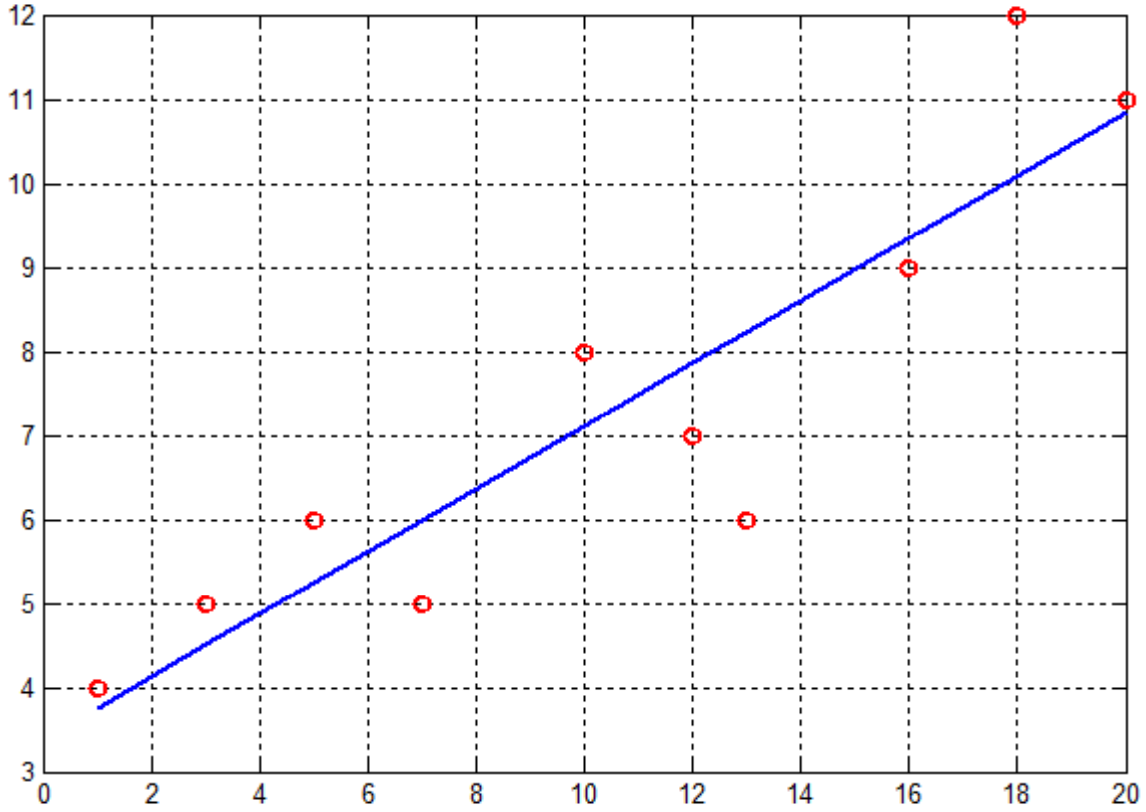
MATLAB implementation with `polyfit`

```
polyfit - Polynomial curve fitting

    This MATLAB function finds the coefficients of a polynomial p(x) of degree n
    that fits the data, p(x(i)) to y(i), in a least squares sense.

    p = polyfit(x,y,n)
    [p,S] = polyfit(x,y,n)
    [p,S,mu] = polyfit(x,y,n)
```

```
x=[1;3;5;7;10;12;13;16;18;20];
y=[4;5;6;5;8;7;6;9;12;11];
p=polyfit(x,y,1)
yapp=p(2)+p(1)*x
plot(x,y,'ro')
grid
hold on
plot(x,yapp)
```



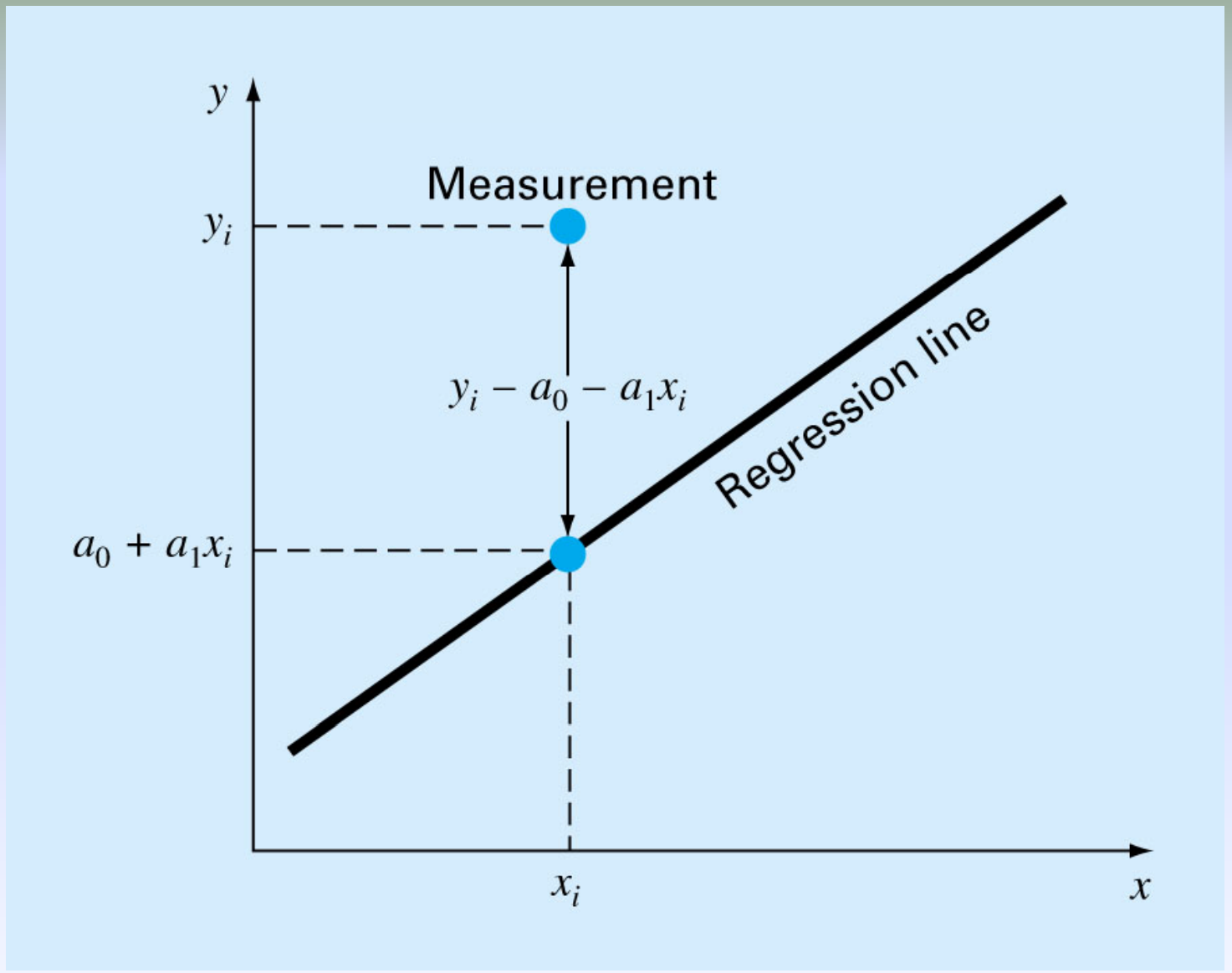


FIGURE 17.3
The residual in linear regression represents the vertical distance between a data point and the straight line.

Error minimization (which one is the best fit?)

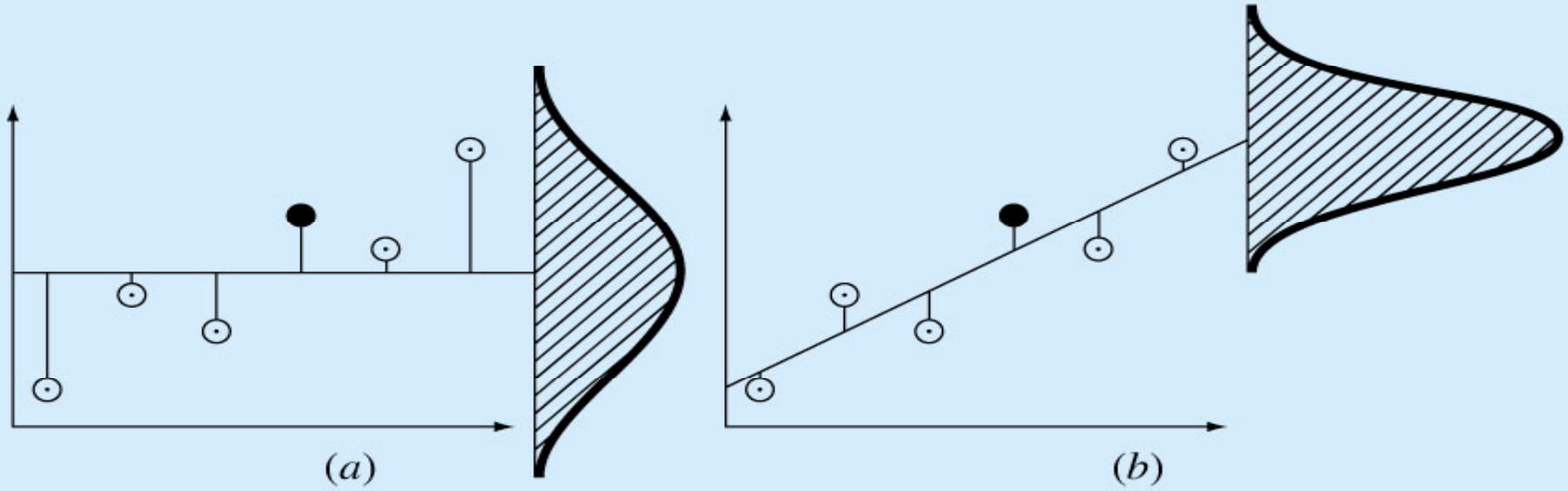


FIGURE 17.4
Regression data showing (a) the spread of the data around the mean of the dependent variable and (*b*) the spread of the data around the best-fit line. The reduction in the spread in going from (*a*) to (*b*), as indicated by the bell-shaped curves at the right, represents the improvement due to linear regression.

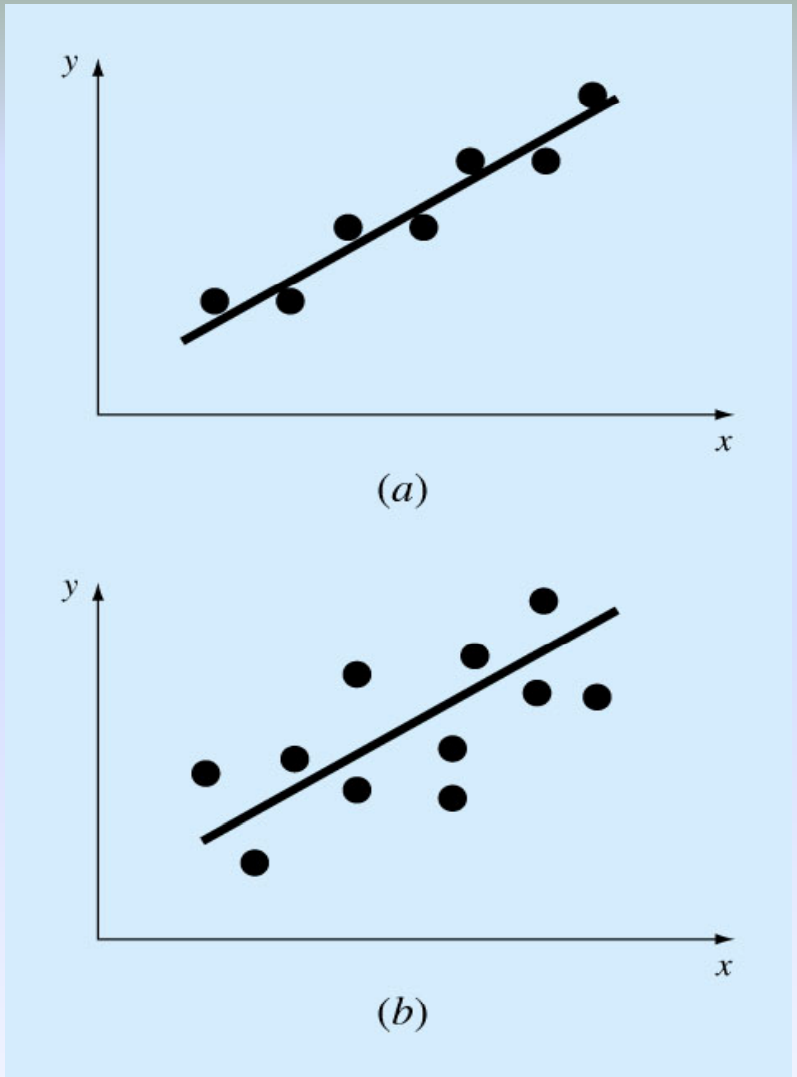


FIGURE 17.5
Examples of linear regression with (*a*) small and (*b*) large residual errors.

"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.
- Usually an r value close to 1 represents a good fit. But be careful and always plot the data points and the regression line together to see what is going on.

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

$$r = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{n\Sigma x_i^2 - (\Sigma x_i)^2}\sqrt{n\Sigma y_i^2 - (\Sigma y_i)^2}}$$

Homework:

- The following pseudocode is an algorithm for linear regression. Implement this algorithm by using MATLAB and try it with the data of the previous example (the example which we used the *polyfit* command).

```
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 + x_i
    sumy = sumy + y_i
    sumxy = sumxy + x_i*y_i
    sumx2 = sumx2 + x_i*x_i
  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 + (y_i - ym)^2
    sr = sr + (y_i - a1*x_i - a0)^2
  END DO
  syx = (sr/(n - 2))^0.5
  r2 = (st - sr)/st

END Regress
```

FIGURE 17.6
Algorithm for linear regression.

Linearization of nonlinear behavior

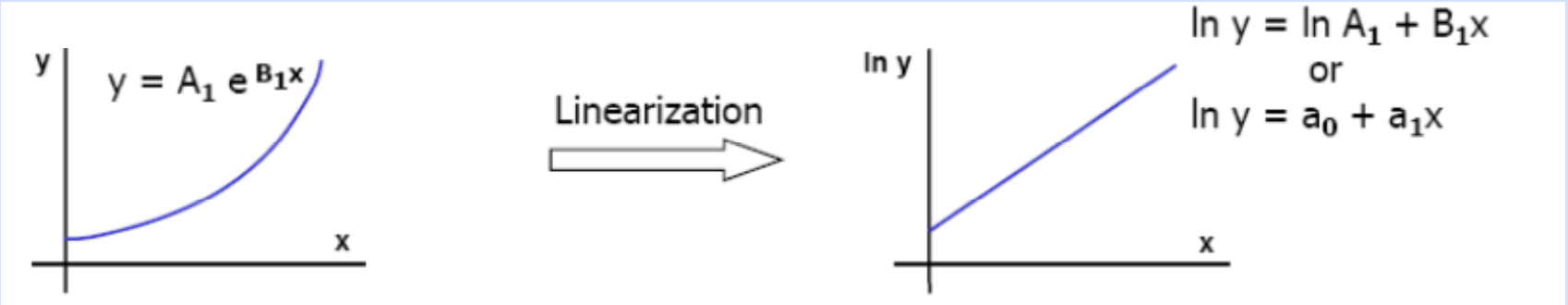
- Linear regression is useful to represent a linear relationship.
- If the relation is nonlinear **either** *another technique can be used* **or** *the data can be transformed so that linear regression can still be used*. The latter technique is frequently used to fit the the following nonlinear equations to a set of data.

$$\text{Exponential equation}: y = A_1 e^{B_1 x}$$

$$\text{Power equation}: \; y = A_2 x^{B_2}$$

$$\text{Saturation - growth rate equation}: y = A_3 x/(B_3 + x)$$

a. EXPONENTIAL EQUATION



Example: Fit *an exponential model* to the following data

x	0.4	0.8	1.2	1.6	2.0	2.3
y	750	1000	1400	2000	2700	3750

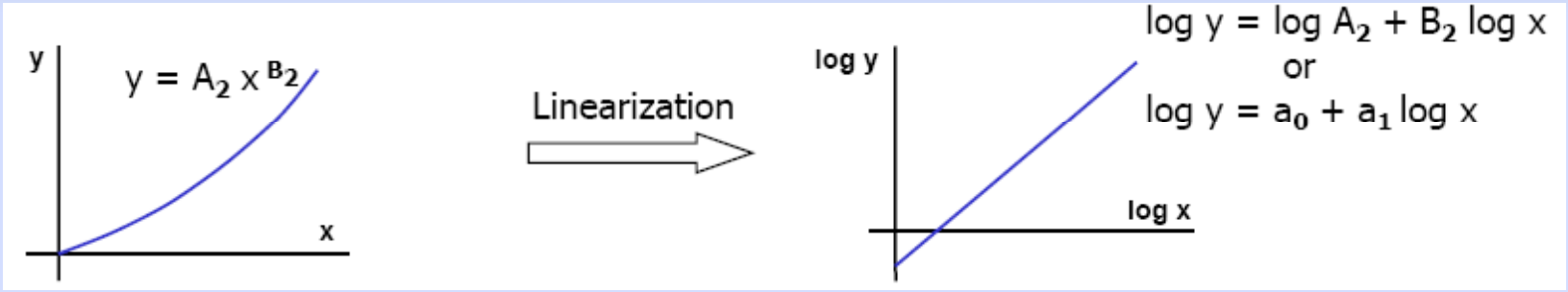
Create the following table

x	0.4	0.8	1.2	1.6	2.0	2.3
ln y	6.62	6.91	7.24	7.60	7.90	8.23

Procedure:

- Fit a straight line to this new data set. Be careful with the notation.
- You can define $z=\ln y$.
- Calculate $a_0=6.25$ and $a_1=0.841$. Straight line is $\ln y=6.25+0.841x$
- Switch back to the original equation. $A_1=e^{a0}=518$, $B_1=a_1=0.841$
- The exponential equation is $y=518\ e^{0.841x}$
- Check this solution with couple of data points.
  - For example $y(1.2)=518\ e^{0.841(1.2)}=1421$
  - $y(2.3)= 518\ e^{0.841(2.3)}=3584$

b. POWER EQUATION



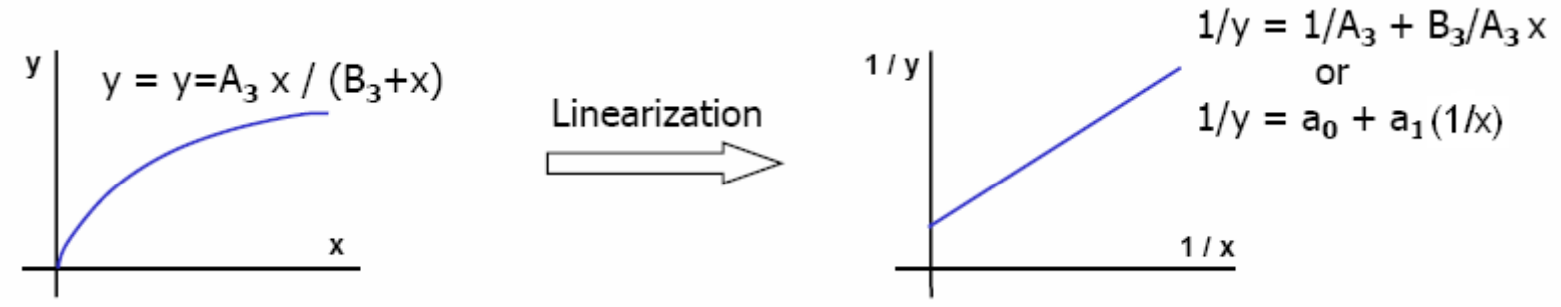
Example: Fit *a power equation* to the following data set

x	2.5	3.5	5	6	7.5	10	12.5	15	17.5	20
y	7	5.5	3.9	3.6	3.1	2.8	2.6	2.4	2.3	2.3

log x	0.398	0.544	0.699	0.778	0.875	1.000	1.097	1.176	1.243	1.301
log y	0.845	0.740	0.591	0.556	0.491	0.447	0.415	0.380	0.362	0.362

- Fit a straight line to this new data set. Be careful with the notation.
- Calculate $a_0 = 1.002$ and $a_1 = -0.53$. Straight line is $\log y = 1.002 - 0.53 \log x$
- Switch back to the original equation. $A_2 = 10^{a_0} = 10.05$, $B_2 = a_1 = -0.53$.
- Therefore the power equation is $y = 10.05\ x^{-0.53}$. Check this solution with couple of data points. For example $y(5) = 10.05 * 5^{-0.53} = 4.28$ or $y(15) = 10.05 * 15^{-0.53} = 2.39$. OK.

c. SATURATION-GROWTH RATE EQUATION



Example: Fit *a saturation-growth rate equation* to the following data set

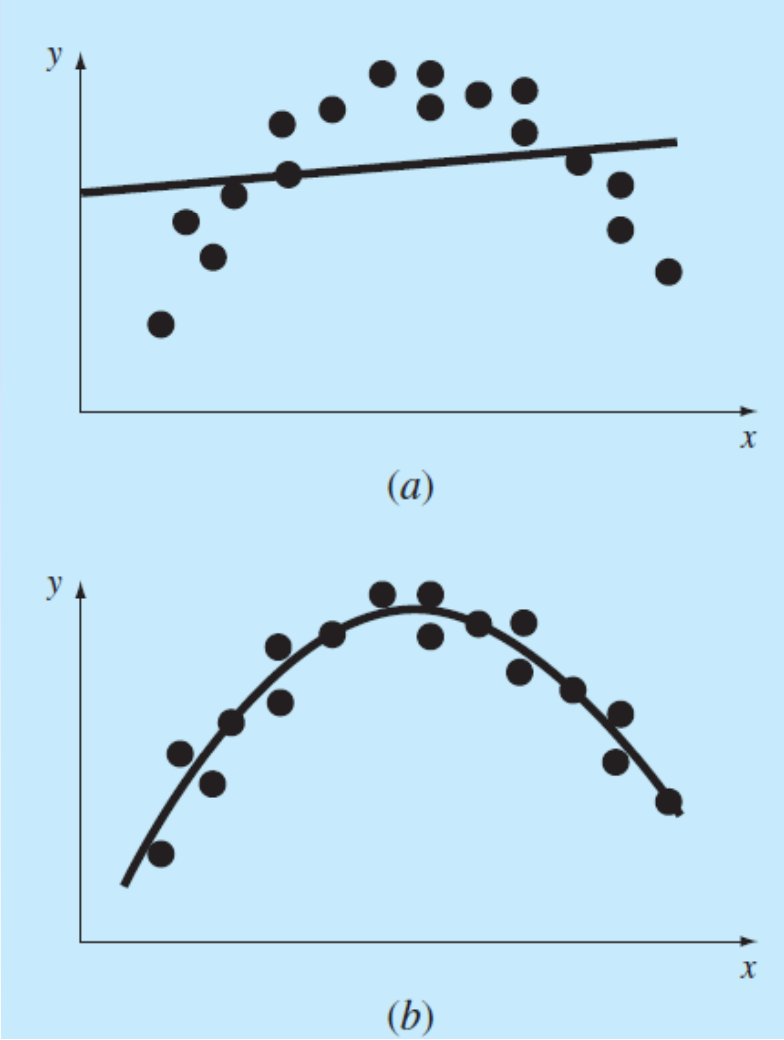
x	0.75	2	2.5	4	6	8	8.5
y	0.8	1.3	1.2	1.6	1.7	1.8	1.7

1 / x	1.333	0.5	0.4	0.25	0.1667	0.125	0.118
1 / y	1.25	0.769	0.833	0.625	0.588	0.556	0.588

- Fit a straight line to this new data set. Be careful with the notation.
- Calculate $a_0 = 0.512$ and $a_1 = 0.562$. Straight line is $1/y = 0.512 + 0.562\ (1/x)$
- Switch back to the original equation. $A_3 = 1/a_0 = 1.953$, $B_3 = a_1 A_3 = 1.097$.
- Therefore the saturation-growth rate equation is $1/y = 1.953\ x/(1.097+x)$. Check this solution with couple of data points. For example $y(2) = 1.953*2/(1.097+2) = 1.26$ OK.

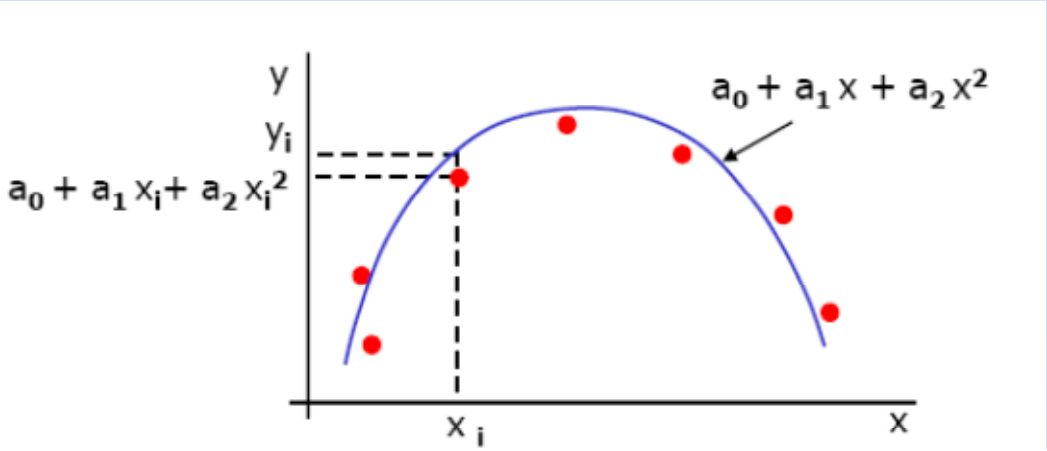
Polynomial Regression

- 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.



Polynomial Regression

- Used to find a best-fit line for a nonlinear behavior.



Example for a second order polynomial regression: $y = a_0 + a_1 x + a_2 x^2 + e$

$e_i = y_i - a_0 - a_1 x_i - a_2 x_i^2$ Error (deviation) for the i^{th} data point

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

To minimize S_r, solve these equations to determine a_0, a_1, and a_2

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum x_i^2 (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

Example:

Polynomial Regression

Problem Statement. Fit a second-order polynomial to the data

x_i	y_i
0	2.1
1	7.7
2	13.6
3	27.2
4	40.9
5	61.1

Solution. From the given data,

$$m = 2 \qquad \sum x_i = 15 \qquad \sum x_i^4 = 979$$

$$n = 6 \qquad \sum y_i = 152.6 \qquad \sum x_i y_i = 585.6$$

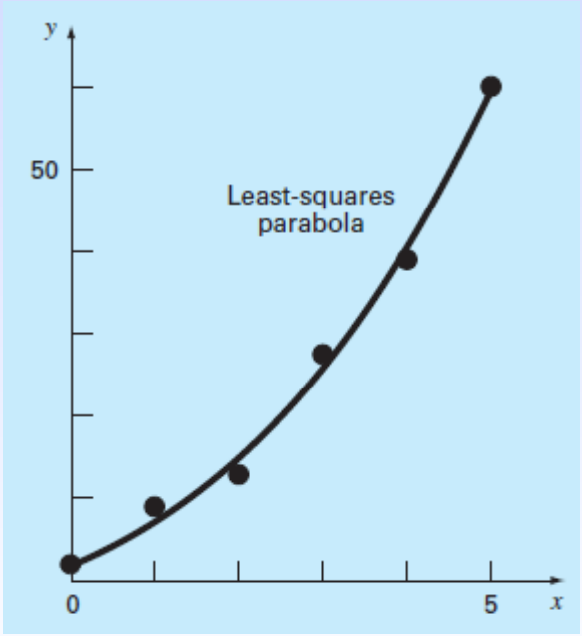
$$\bar{x} = 2.5 \qquad \sum x_i^2 = 55 \qquad \sum x_i^2 y_i = 2488.8$$

$$\bar{y} = 25.433 \qquad \sum x_i^3 = 225$$

Therefore, the simultaneous linear equations are

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

$a_0 = 2.47857$ $a_1 = 2.35929$, and $a_2 = 1.86071$



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

Homework:

- Try the following pseudocode by using MATLAB with some arbitrary examples and compare the results on the same examples with `polyfit` command.

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 $a_0, a_1, a_2, \ldots, a_m$, using an elimination method.
Step 6: Print out the coefficients.

```
DOFOR i = 1, order + 1
  DOFOR j = 1, i
    k = i + j − 2
    sum = 0
    DOFOR ℓ = 1, n
      sum = sum + x_ℓ^k
    END DO
    a_i,j = sum
    a_j,i = sum
  END DO
  sum = 0
  DOFOR ℓ = 1, n
    sum = sum + y_ℓ · x_ℓ^(i−1)
  END DO
  a_i,order+2 = sum
END DO
```

FIGURE 17.13
Pseudocode to assemble the elements of the normal equations for polynomial regression.

MULTIPLE LINEAR REGRESSION

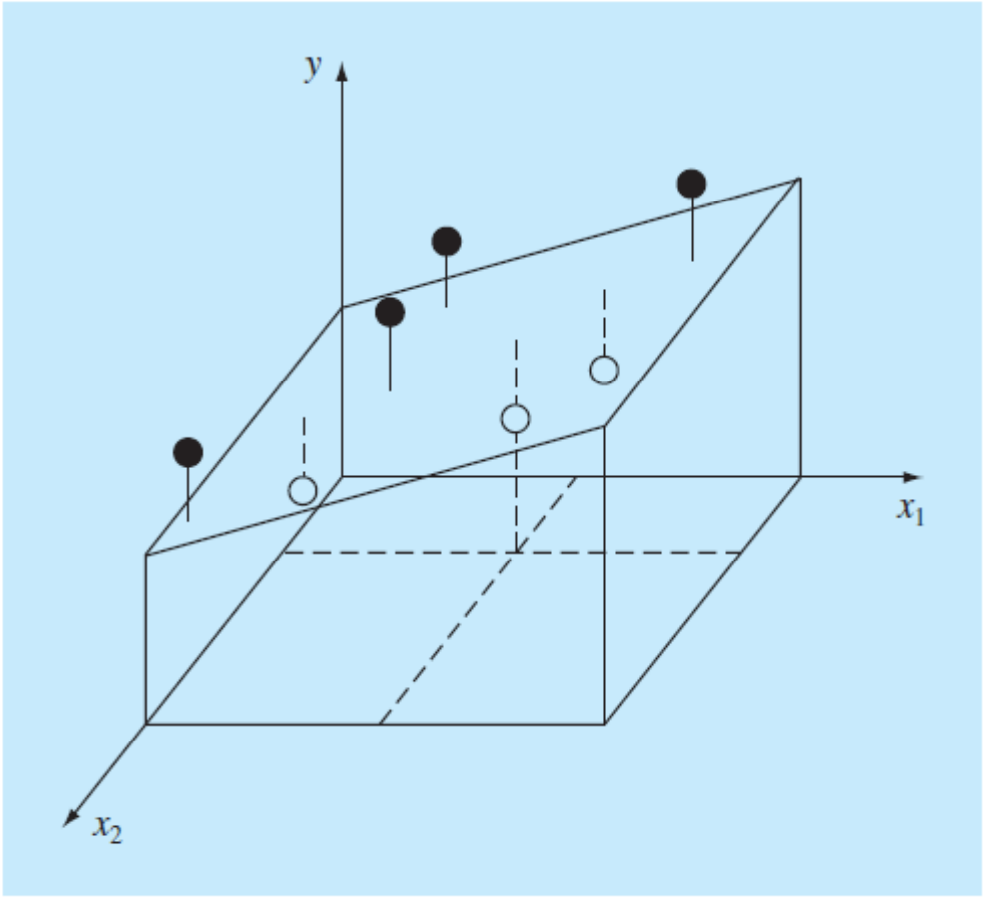
- A useful extension of linear regression is the case where y is a linear function of two or more independent variables. For example, y might be a linear function of 2 variables: x_1 and x_2, as in

$$y = a_0 + a_1 x_1 + a_2 x_2 + e$$

- For this two-dimensional case, the regression "line" becomes a "plane".

MULTIPLE LINEAR REGRESSION

FIGURE 17.14
Graphical depiction of multiple linear regression where y is a linear function of x_1 and x_2.



MULTIPLE LINEAR REGRESSION

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

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

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

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

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

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

Example:

Problem Statement. The following data were 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 these 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 these data were derived.

MULTIPLE LINEAR REGRESSION

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

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

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

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:

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

MULTIPLE LINEAR REGRESSION

FIGURE 17.15

Pseudocode to assemble the elements of the normal equations for multiple regression. Note that aside from storing the independent variables in $x_{1,i}$, $x_{2,i}$, etc., 1's must be stored in $x_{0,i}$ for this algorithm to work.

```
DOFOR i = 1, order + 1
  DOFOR j = 1, i
    sum = 0
    DOFOR ℓ = 1, n
      sum = sum + x_{i-1,ℓ} · x_{j-1,ℓ}
    END DO
    a_{i,j} = sum
    a_{j,i} = sum
  END DO
  sum = 0
  DOFOR ℓ = 1, n
    sum = sum + y_ℓ · x_{i-1,ℓ}
  END DO
  a_{i,order+2} = sum
END DO
```

General Linear Least Squares

- In the preceding pages, we have introduced three types of regression:
  - simple linear,
  - polynomial,
  - multiple linear.
- In fact, all three belong to the following general linear least-squares model:

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

where $z_0, z_1, \ldots, z_m$ are $m + 1$ basis functions. It can easily be seen how simple and multiple linear regression fall within this model—that is, $z_0 = 1$, $z_1 = x_1$, $z_2 = x_2$, $\ldots$, $z_m = x_m$. Further, polynomial regression is also included if the basis functions are simple monomials as in $z_0 = x^0 = 1$, $z_1 = x$, $z_2 = x^2$, $\ldots$, $z_m = x^m$.

General Linear Least Squares

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

can be expressed in vector-matrix notations:

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

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

$\{Y\}$ – observed valued 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