Chapter 13
Curve Fitting and Correlation

This chapter will be concerned primarily with two separate but closely interrelated processes: (1) the fitting of experimental data to mathematical forms that describe their behavior and (2) the correlation between different experimental data to assess how closely different variables are interdependent.

The fitting of experimental data to a mathematical equation is called *regression.* Regression may be characterized by different adjectives according to the mathematical form being used for the fit and the number of variables. For example, *linear regression* involves using a straight-line or linear equation for the fit. As another example, *Multiple regression* involves a function of more than one independent variable.

Linear Regression

Assume *n* points, with each point having values of both an independent variable *x* and a dependent variable *y*.

The values of x are $x_1, x_2, x_3, ..., x_n$.

The values of y are $y_1, y_2, y_3, ..., y_n$.

A best-fitting straight line equation

will have the form

$$y = a_1 x + a_0$$

Preliminary Computations

$$\bar{x} = \text{sample mean of the } x \text{ values} = \frac{1}{n}\sum_{k=0}^{n} x_k$$

$$\bar{y} = \text{sample mean of the } y \text{ values} = \frac{1}{n}\sum_{k=0}^{n} y_k$$

$$\overline{x^2} = \text{sample mean-square of the } x \text{ values} = \frac{1}{n}\sum_{k=1}^{n} x_k^2$$

$$\overline{xy} = \text{sample mean of the product } xy = \frac{1}{n}\sum_{k=1}^{n} x_k y_k$$

Best-Fitting Straight Line

$$a_1 = \frac{\overline{xy} - \bar{x}\ \bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$a_0 = \frac{\overline{x^2}\ \bar{y} - \bar{x}\ \overline{xy}}{\overline{x^2} - \bar{x}^2}$$

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

$$y = a_1 x + a_0$$

Example 13-1. Find best fitting straight line equation for the data shown below.

x	0	1	2	3	4	5	6	7	8	9
y	4.00	6.10	8.30	9.90	12.40	14.30	15.70	17.40	19.80	22.30

$$\bar{x} = \frac{1}{10}\sum_{k=1}^{10} x_k = \frac{0+1+2+3+4+5+6+7+8+9}{10} = \frac{45}{10} = 4.50$$

$$\bar{y} = \frac{1}{10}\sum_{k=1}^{10} y_k = \frac{4+6.1+8.3+9.9+12.4+14.3+15.7+17.4+19.8+22.3}{10}$$

$$= \frac{130.2}{10} = 13.02$$

Example 13-1. Continuation.

$$\overline{x^2} = \frac{1}{10}\sum_{k=1}^{10} x_k^2$$

$$= \frac{(0)^2 + (1)^2 + (2)^2 + (3)^2 + (4)^2 + (5)^2 + (6)^2 + (7)^2 + (8)^2 + (9)^2}{10}$$

$$= \frac{285}{10} = 28.50$$

$$\overline{xy} = \frac{1}{10}\sum_{k=1}^{10} x_k y_k$$

$$= \frac{0 + 6.1 + 16.6 + 29.7 + 49.6 + 71.5 + 94.2 + 121.8 + 158.4 + 200.7}{10}$$

$$= \frac{748.6}{10} = 74.86$$

Example 13-1. Continuation.

$$a_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{74.86 - (4.50)(13.02)}{28.50 - (4.50)^2}$$

$$= \frac{16.27}{8.250} = 1.9721$$

$$a_0 = \bar{y} - a_1\bar{x} = 13.02 - 1.972 \times 4.50 = 4.1455$$

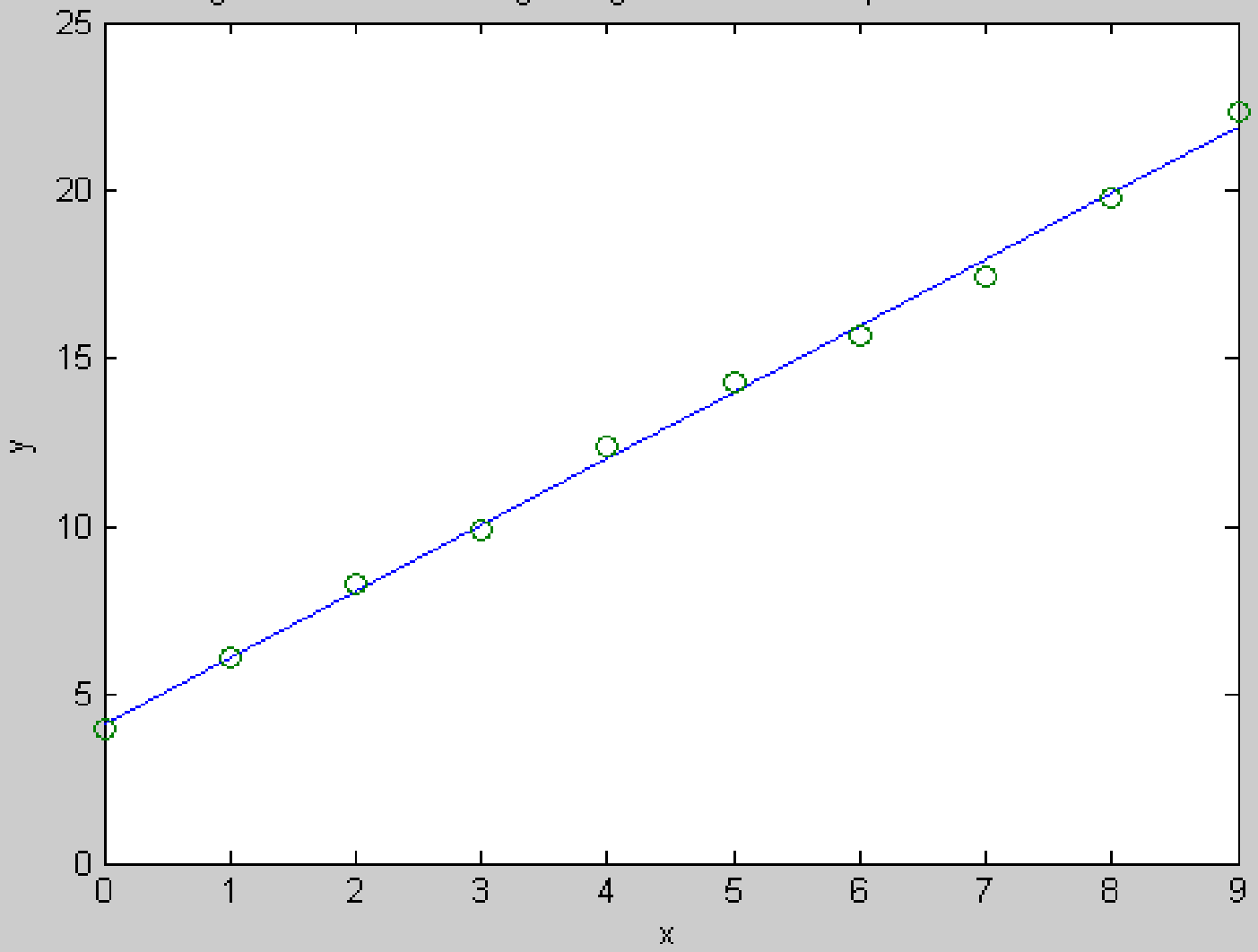
$$y = 1.9721x + 4.1455$$

Example 13-1. Continuation.

```
>> x = 0:9;
>> yapp = 1.9721*x + 4.1455;
>> y = [the 10 values of y];
>> plot(x, yapp, x, y, 'o')
```

The best-fit plot and the actual points are shown on the next slide.

Figure 13-1. Best fitting straight-line for Examples 13-1 and 13-2.



MATLAB General Polynomial Fit

The values of x are $x_1, x_2, x_3, ..., x_n$.

The values of y are $y_1, y_2, y_3, ..., y_n$.

The polynomial fit is to be of the form

$$y = p(x) = a_m x^m + a_{m-1} x^{m-1} + ... + a_1 x + a_0$$

```
>> x = [x1 x2 x3.......xn];
>> y = [y1 y2 y3......yn];
>> p = polyfit(x, y, m)
>> yapp = polyval(p, x)
>> plot(x, yapp, x, y, 'o')
```

Example 13-2. Rework Example 13-1 using MATLAB.

```
>> x = 0:9;
>> y = [the 10 values of y];
>> p = polyfit(x, y, 1)
p =
   1.9721    4.1455
```

These are the same values obtained manually in Example 13-1.

Example 13-3. For data of previous two examples, obtain a 2nd degree fit.

Assume that the vectors x and y are still in memory.

```
>> p = polyfit(x, y, 2)
p =
   0.0011    1.9619    4.1591
>> yapp2 = polyval(p, x);
>> plot(x, yapp2, x, y, 'o')
```

The results are shown on the next slide.

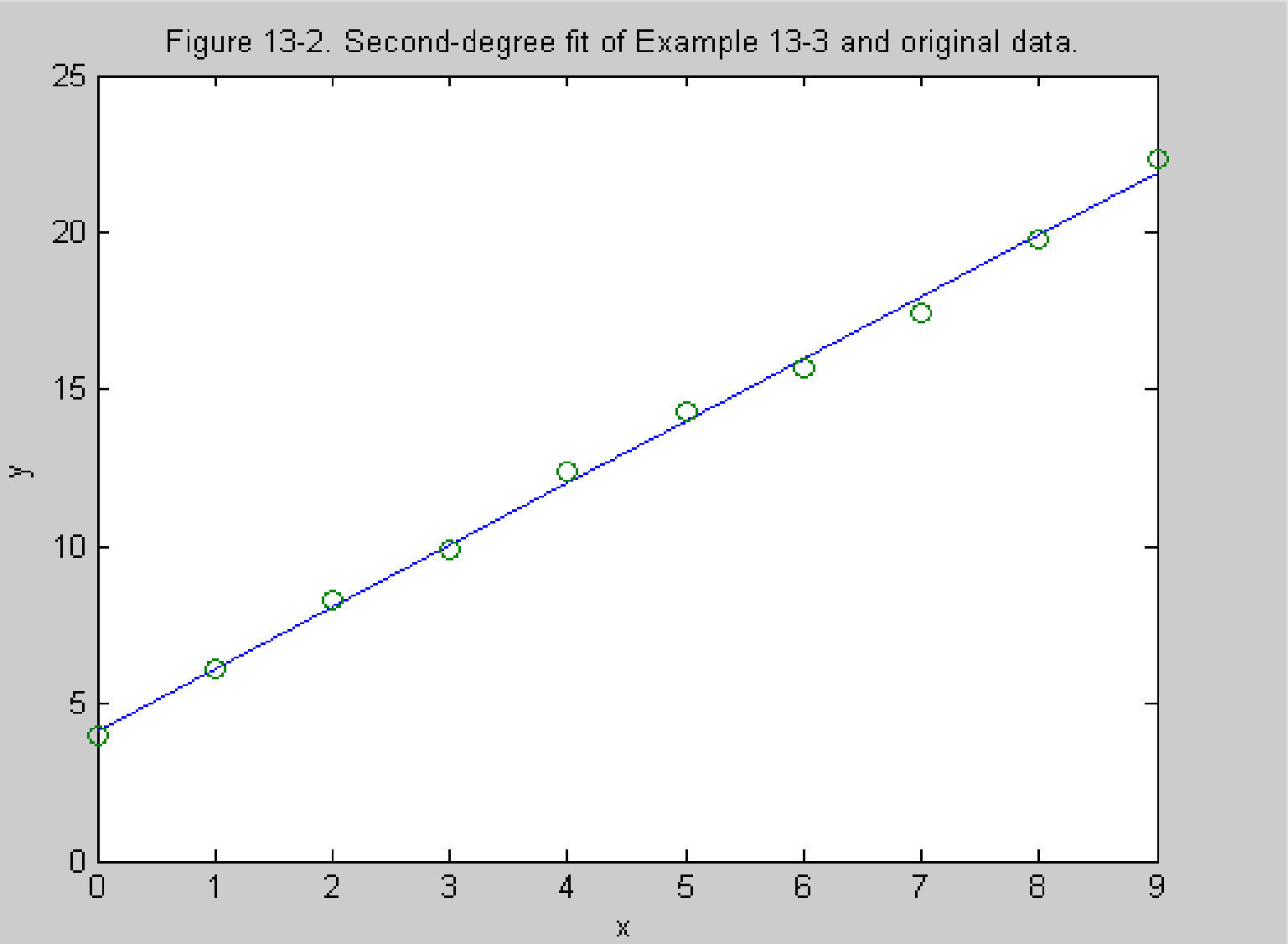


Figure 13-2. Second-degree fit of Example 13-3 and original data.
25
20
15
10
5
0
y
0
1
2
3
4
5
6
7
8
9
x

Example 13-4. Determine several polynomial fits for the function below.

$$y = \sin \pi t \qquad \text{for} \qquad -1 \le t \le 1$$

```
>>  t = -1:0.05:1;
>> y = sin(pi*t);
>> plot(t, y)
```

A plot of the function is shown on the next slide.

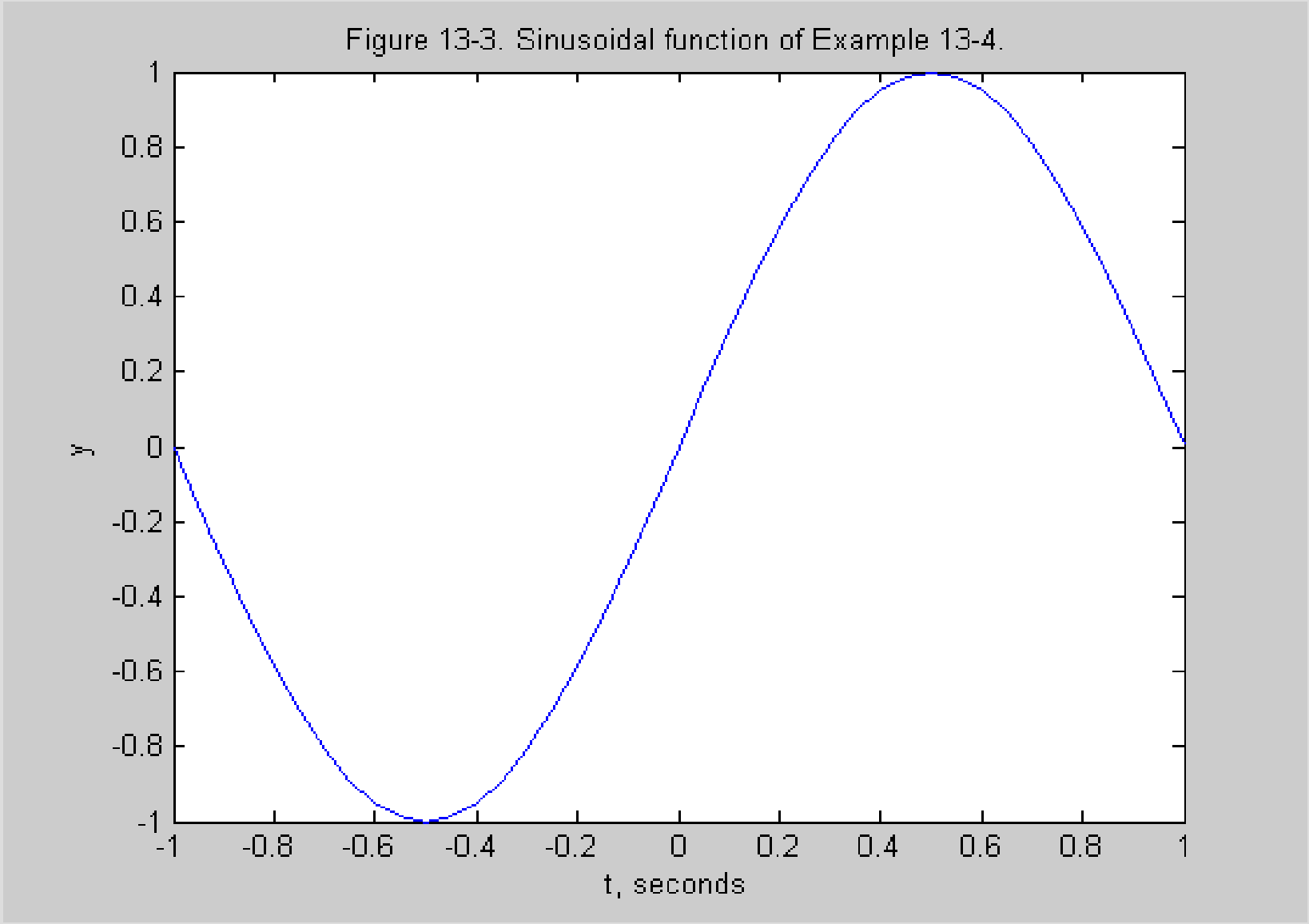


Figure 13-3. Sinusoidal function of Example 13-4.
y
t, seconds
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

Example 13-4. Continuation.

(a) m = 1

```
>> p1 = polyfit(t, y, 1)
p1 =
   0.8854    0.0000
>> yapp1 = polyval(p1, t);
>> plot(t, yapp1, t, y, 'o')
```

The results are shown on the next slide.

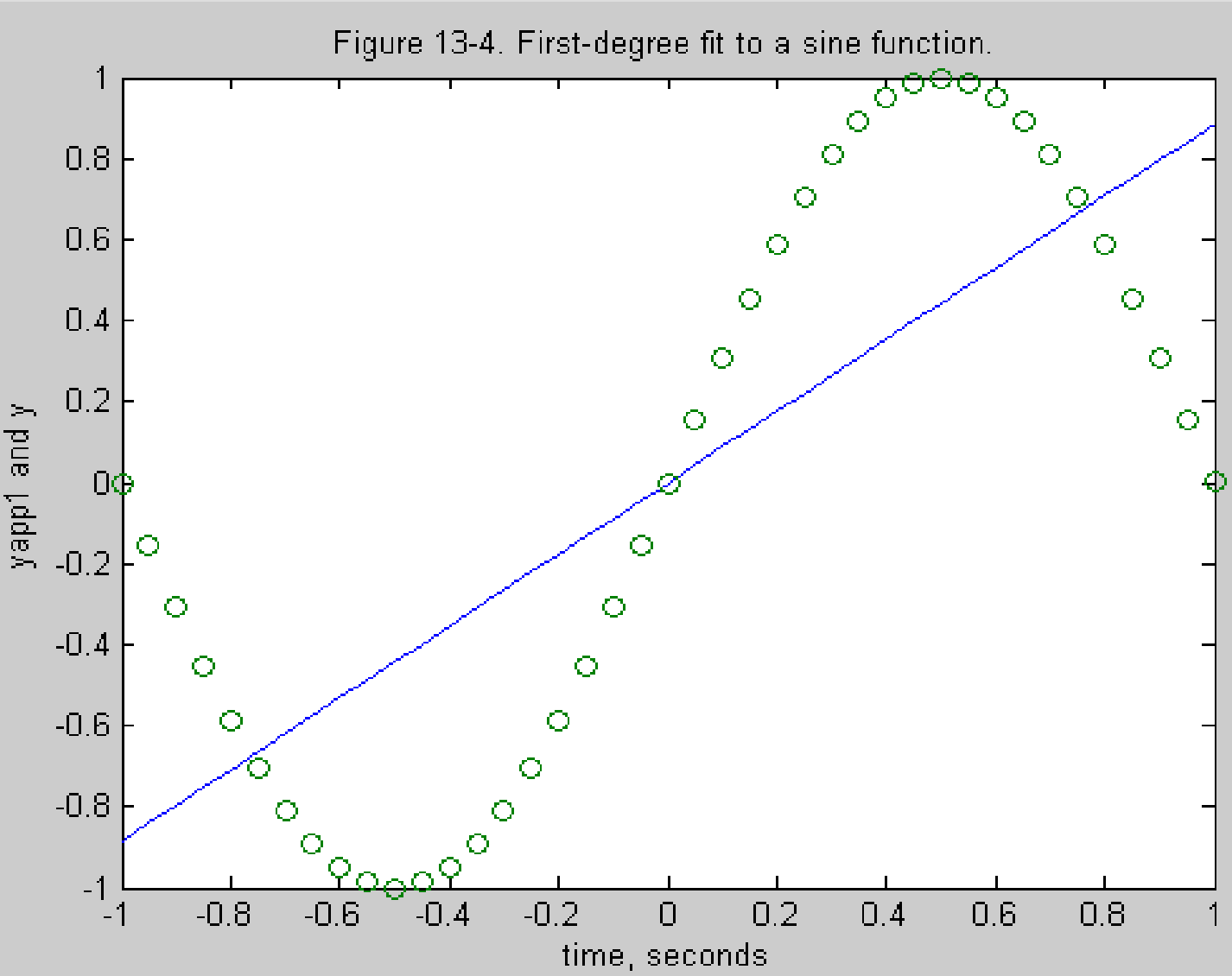


Figure 13-4. First-degree fit to a sine function.
yappl and y
time, seconds
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

Example 13-4. Continuation.

(b) m = 2

```
>> p2 = polyfit(t, y, 2)
p2 =
    0.0000    0.8854   -0.0000
```

The polynomial is the same as for m = 1. This is due to the fact that the sine function is an odd function and the coefficients of the terms with even degrees are zero.

Example 13-4. Continuation.

(c) m = 3

```
>> p3 = polyfit(t, y, 3)
p3 =
   -2.8139   -0.0000    2.6568    0.0000
>> yapp3 = polyval(p3, t);
>> plot(t, yapp3, t, y, 'o')
```

The results are shown on the next slide. A fit for m = 4 would be the same as for m = 3.

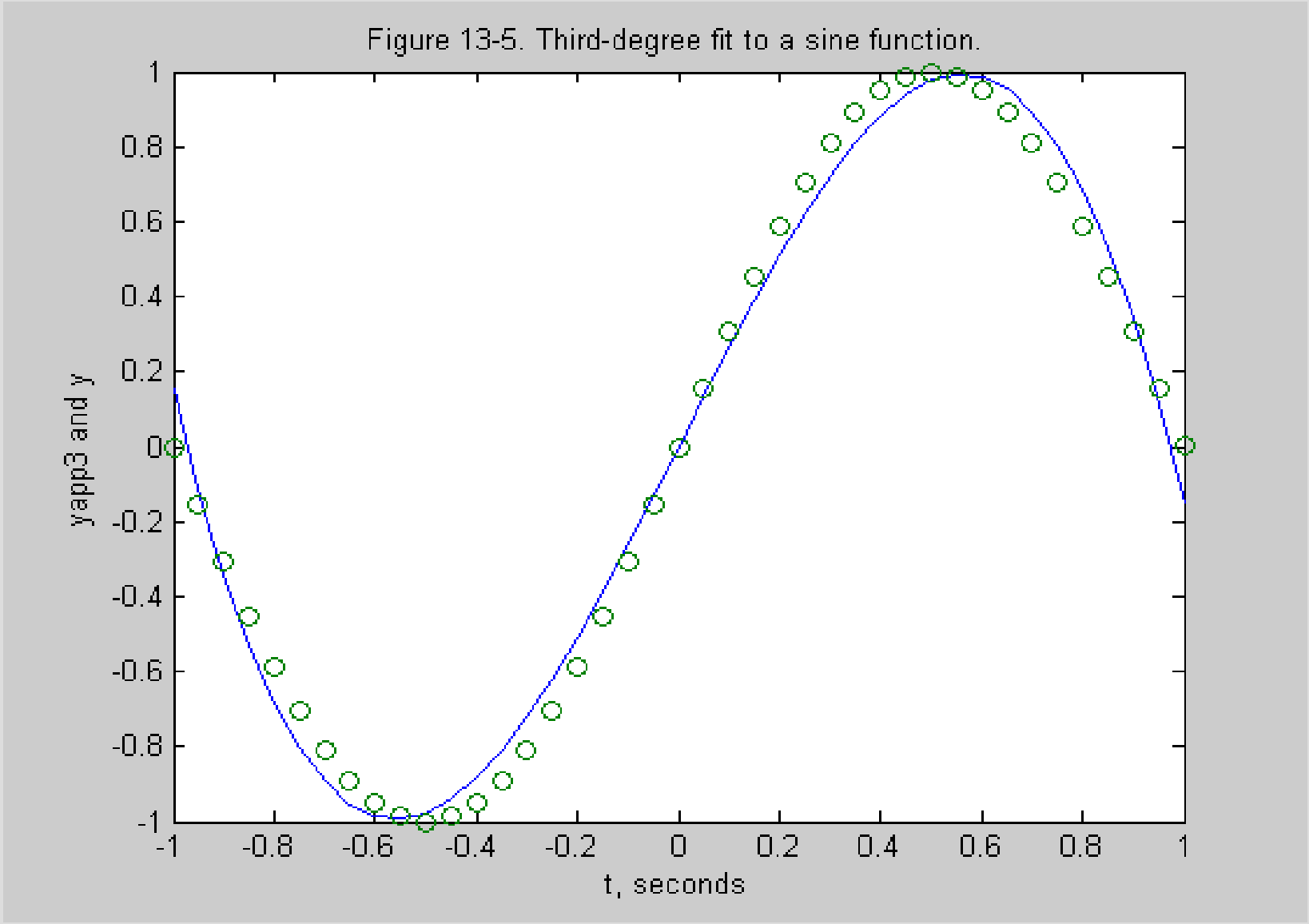


Figure 13-5. Third-degree fit to a sine function.
yapp3 and y
t, seconds
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

Example 13-5. Continuation.

```
m = 5
>> p5 = polyfit(t, y, 5)
p5 =
   1.6982    0.0000   -4.7880   -0.0000
  3.0990    0.0000
>> yapp5 = polyval(p5, t);
>> plot(t, yapp5, t, y, 'o')
```

The results are shown on the next slide.

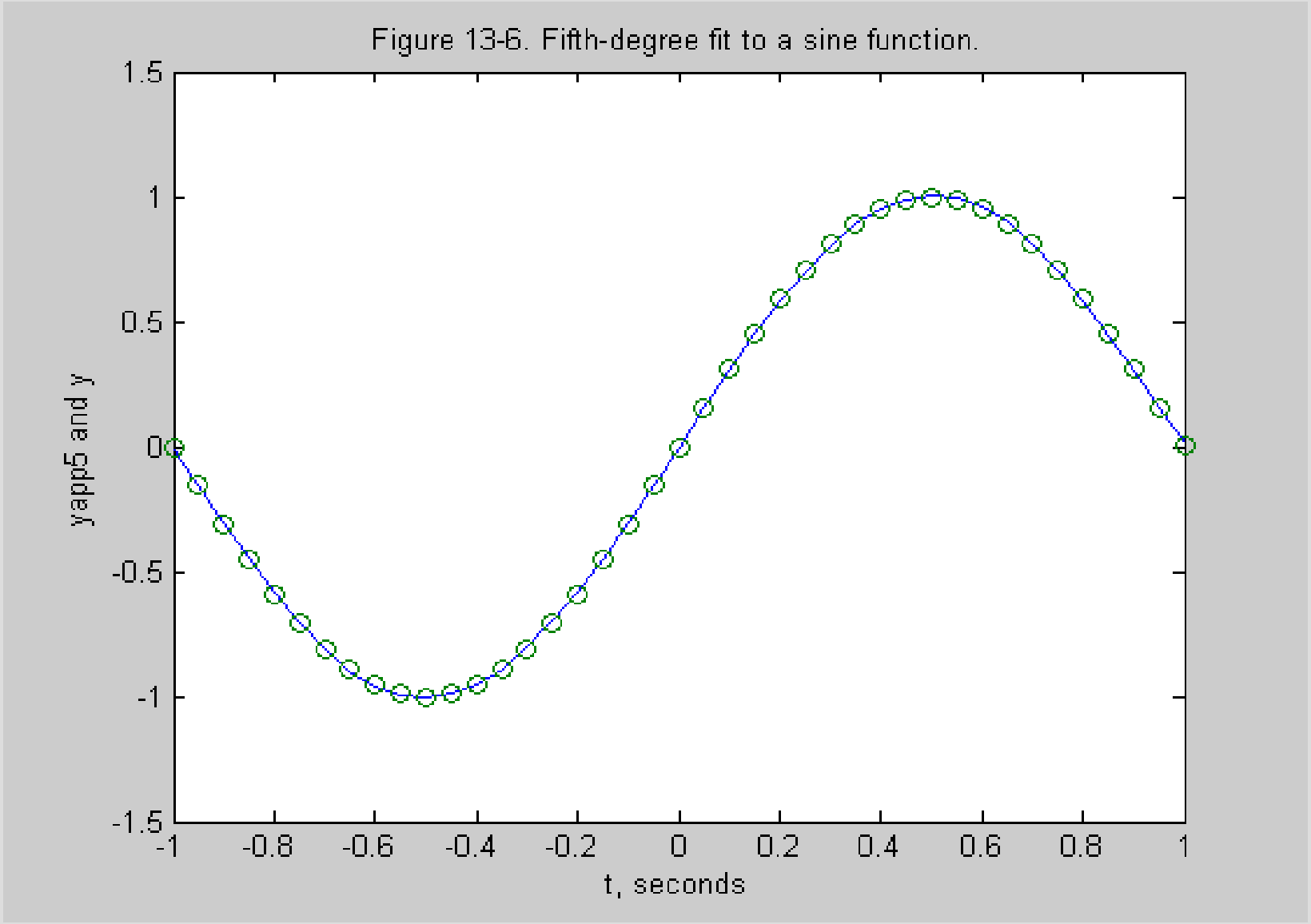


Figure 13-6. Fifth-degree fit to a sine function.
1.5
1
0.5
0
-0.5
-1
-1.5
yapp5 and y
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
t, seconds

Example 13-5. For data below, obtain a 2nd degree fit for the temperature *T* as a function of the distance *x*.

x(ft)	0	1	2	3	4	5
T(deg F)	71	76	86	100	118	140

```
>> x = 0:5;
>> T = [71 76 86 100 118 140];
>> p = polyfit(x,T,2)
p =
    2.0893    3.4107   70.8214
```

Example 13-5. Continuation.

The equation is

$$T = 2.0893x^2 + 3.4107x + 70.8214$$

```
>> x1 = 0:0.1:5;
>> T1 = polyval(p, x1);
>> plot(x1, T1, x, T, 'o')
```

The results are shown on the next slide.

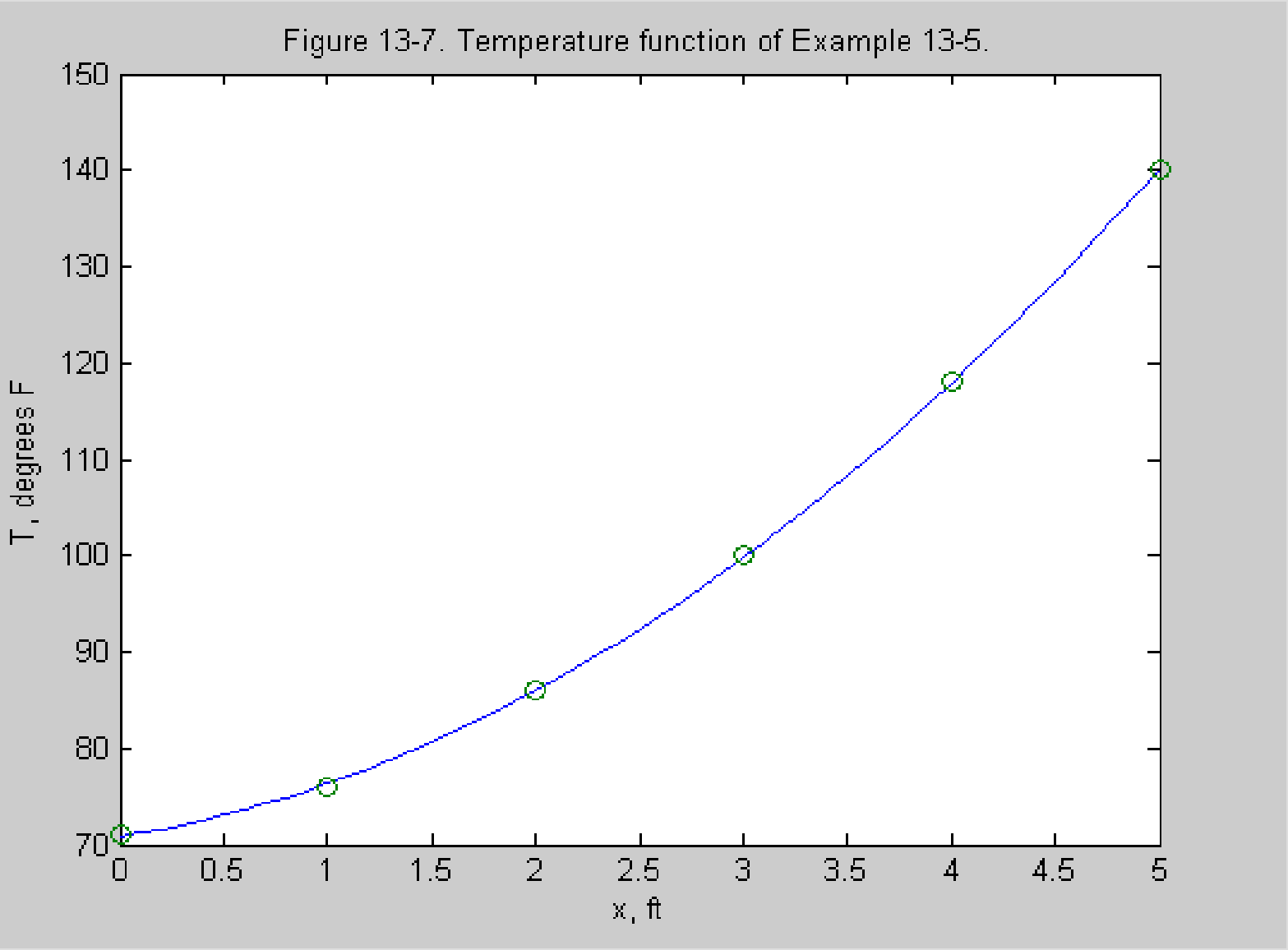


Figure 13-7. Temperature function of Example 13-5.
T, degrees F
x, ft
150
140
130
120
110
100
90
80
70
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

Multiple Linear Regression

Assume m independent variables

$$x_1,\ x_2,.....x_m$$

Assume a dependent variable y that is to be considered as a linear function of the m independent variables.

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m$$

Multiple Regression (Continuation)

Assume that there are k values of each of the m variables. For x_1, we have

$$x_{11}, x_{12}, x_{13}, \ldots\ldots, x_{1k}$$

Similar terms apply for all other variables.

For the mth variable, we have

$$x_{m1}, x_{m2}, x_{m3}, \ldots\ldots, x_{mk}$$

MATLAB Procedure for Linear Regression

1. Form m column vectors each of length k representing the independent variables.

```
>> x1 = [x11 x12 x13.....x1k]';
>> x2 = [x21 x22 x23.....x2k]';
.
.
>> xm = [xm1 xm2 xm3.....xmk]';
```

MATLAB Procedure (Continuation)

2. Form a column vector of length k representing the dependent variable y.

```
>> y = [y1 y2 y3.....yk]';
```

3. Form a rectangular matrix X of size k by m+1 as follows:

```
>> X= [ones(size(x1)) x1 x2 ......xm];
```

4. Determine a column vector a of length m+1 by the command that follows:

```
>> a = X\y
```

MATLAB Procedure (Continuation)

5. The best-fit linear multiple regression formula is then given by

```
>> Y = X*a;
```

6. The maximum difference between the actual data and the formula is

```
>> Error_Maximum = max(abs(Y-y))
```

Correlation

Cross-Correlation

$$\text{corr}(x, y) = E(xy) = \overline{xy}$$

Covariance

$$\text{cov}(x, y) = E\ \ (x - \bar{x})(y - \bar{y})$$

$$= \text{corr}(x, y) - (\bar{x})(\bar{y})$$

$$= \overline{xy} - (\bar{x})(\bar{y})$$

Correlation Coefficient

$$C(x,y) = \frac{E\ \ (x-\bar{x})(y-\bar{y})}{\sigma_x \sigma_y}$$

$$= \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{cov}(x,x)\,\mathrm{cov}(y,y)}}$$

Implications of Correlation Coefficient

1. If $C(x, y) = 1$, the two variables are totally correlated in a **positive** sense.

2. If $C(x, y) = -1$, the two variables are totally correlated in a **negative** sense.

3. If $C(x, y) = 0$, the two variables are said to be **uncorrelated**.

One Final Note

Correlation does **not** necessarily imply **causation**!