# Exercises

**3.4.1** Use the least–squares matrix approach to fit a linear model $y = a + mx$ to a randomly generated set of data with 4 points. Show by examples that this gives the same results as the formulas given in Chapter 2.

**3.4.2** In this section we found the parameters of a model by forming a system of linear equations and then finding a least-squares solution. Explain why we *cannot* use this approach to find the parameters for a model of the form $y = a\cos(bx)$.

**3.4.3** Fit different polynomial models to the data in the table below and select the "best" one. Explain how you determined which one is "best."

| **x** | 1.4 | 2.4 | 7.1 | 13.8 | 34.2 | 109.3 | 134 |
|---|---|---|---|---|---|---|---|
| **y** | 2.7 | 2.27 | 3.31 | 3.39 | 3.81 | 4.88 | 4.62 |

**3.4.4** Calculate the $R^2$ value for the model fit to deer population data in Example 3.4.5.

**3.4.5** The sin and cos terms in the deer population model in Example 3.4.5 were included to capture the oscillating pattern of the data. One might wonder if both terms are really necessary. Fit a model of the form $y = a + bx + c\cos(\pi x/5)$ and then fit a model of the form $y = a + bx + d\sin(\pi x/5)$ to the data. Calculate the $R^2$ value for each and compare these revised models to the original model. Does the inclusion of both sin and cos terms yield a "significantly" better model?

**3.4.6** Consider the problem of fitting a $2^{nd}$ polynomial to a set of 4 data points discussed in Example 3.4.2. Add the requirement that $y(0) = c_0$ where $c_0$ is some specified value (in other words we are specifying the $y$-intercept of the model). Design a spreadsheet to find a least-squares $2^{nd}$ degree polynomial model where the user can input 4 data points and specify the value of $c_0$.

**3.4.7** The amount of a radioactive substance remaining after time $t$, $y(t)$, is described by the exponential model $y(t) = Ce^{-kt}$ where $C$ is the initial amount (the amount at time $t = 0$) and $k$ is a constant. Suppose two radioactive substances $A$ and $B$ have constants $k_A = 0.03$ and $k_B = 0.05$. A mixture of these two substances contains $C_A$ grams of $A$ and $C_B$ grams of $B$ at time $t = 0$, both of which are unknown. The total amount of the mixture at time $t$ is modeled by

$$y(t) = C_A e^{-0.03t} + C_B e^{-0.05t} \tag{3.4}$$

A researcher measures the total amount of the mixture at several times and records the data in the table below. Estimate the values of $C_A$ and $C_B$ by fitting a least-squares model of the form (3.4) to the data.

| **Time** | 5 | 6 | 7 | 8 |
|---|---|---|---|---|
| **Amount** | 8.8 | 8.6 | 8.2 | 7.9 |

**3.4.8** One useful property of polynomials is that they are easy to differentiate and integrate. Suppose a researcher observes a particle moving in a straight line and measures the particle's velocity relative to its starting position at several points in time as shown in the table below.

| **t** | 0 | 1.2 | 2.5 | 3.2 | 4.6 | 5.4 | 6.3 | 7.3 |
|---|---|---|---|---|---|---|---|---|
| **v(t)** | 0 | 3.6 | 5.5 | 7.4 | 6.7 | 5.8 | 3.5 | 0 |

a. Fit several polynomial models to the data and choose the one that "best" models the velocity. (**Suggestion**: When choosing the "best" model, don't rely strictly on the $R^2$ value. Put a large emphasis on simplicity.)

b. The acceleration of the particle at time $t$ is $a(t) = v'(t)$. Use your model in part a. to estimate $a(1.8)$.

c. The total distance traveled over the time interval $[a, b]$ is

$$\text{Total distance travled} = \int_a^b |v(t)|\, dt.$$

Use your model in part a. to estimate the total distance the particle traveled over the interval $[0, 7.3]$.

**3.4.9** Consider a refinement to the deer population model in Example 3.4.5. Note that as time increases, the difference between a high point and the next low point (the amplitude) tends to decrease. Our original model did not take this into account.

a. For the years 1947, 1957, 1965, and 1977, calculate the amplitude by subtracting the population in the next year.

b. Create a graph of amplitude vs. year using the data in part a.

c. Fit an exponential model, $g(x) = me^{kx}$, to the data in part a.

d. Fit a model of the form

$$y = a + bx + c\, g(x) * \cos\left(\frac{\pi x}{5}\right) + d\, g(x) * \sin\left(\frac{\pi x}{5}\right)$$

to the original data. Does this model seem to fit the data any better than the original one?

e. Calculate the $R^2$ value for this refined model. How does this compare to the original model?

**3.4.10** The table below contains the temperatures over one day in Seward, NE starting at midnight.

| **Hour** | 0 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 | 22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Temp** | 44.5 | 45.3 | 52.6 | 60.4 | 70.2 | 75.9 | 79.8 | 79.1 | 72.8 | 63.5 | 52.5 | 44.6 |

Our goal is to predict the temperature $y$ at time $x$. Notice that the temperature is periodic, so we will use a model of the form $y = a\sin(bx + c) + d$. Follow these steps to design a spreadsheet to implement a simple "sine regression" algorithm for fitting a model of this type to the data:

a. Enter the data in a spreadsheet, create a graph of the data, and designate cells to hold the values of $a$, $b$, $c$, and $d$.

b. The period of the function $y = a\sin(bx + c) + d$ is $\frac{2\pi}{b}$. Estimate the period of the data and use this to estimate the value of $b$.

c. Initially, let $c = 0$.

d. To find the values of $a$ and $d$, create a graph of $y$ vs. $\sin(bx + c)$ and fit a linear trendline to this transformed data. Display the $R^2$ value. The slope is the value of $a$ and the $y$-intercept is the value of $d$. Use the functions **SLOPE** and **INTERCEPT**to calculate the values of $a$ and $d$, respectively.

e. Create a scroll bar to vary the value of $b$ between -1 and +1 and another scroll bar to vary the value of $c$ between -2 and +2 (see appendix A.4 for more information on scroll bars).

f. Use the scroll bars to find values of $b$ and $c$ that maximize the $R^2$ value.

g. Graph the model on top of the original data. How we does the model fit the data?

## 3.5 Multiple Regression

In previous sections we have discussed predicting the value of one response variable $y$ with one predictor variable $x$. In this section we will discuss using two or more predictor variables $x_1, x_2, \ldots, x_n$. This topic is called **Multiple Regression**.

Consider the problem of predicting the selling price of a house. The selling price is affected by many factors including the age of the house, living area, number of bedrooms, etc. Table 3.5 lists the selling price, living area (in ft$^2$), acres of land, and the number of bedrooms of 10 homes in a neighborhood.

### Example 3.5.1 Single Predictor Variable

Consider the graphs of Selling Price vs. Area and Selling Price vs. Acres as shown in Figure 3.41 along with the linear regression equation for each.

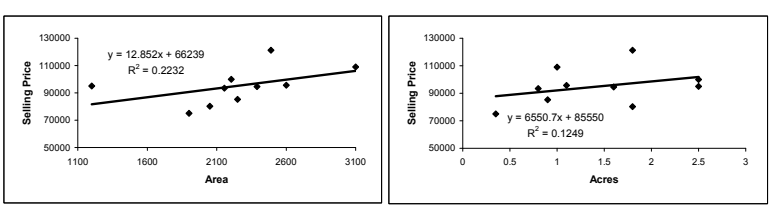


**Figure 3.41**