# Chapter 11

# Linear Statistical Models



## 11.1 The Method of Least Squares

*When each observation from an experiment is a pair of numbers, it is often important to try to predict one of the numbers from the other. Least squares is a method for constructing a predictor of one of the variables from the other by making use of a sample of observed pairs.*

### Fitting a Straight Line

**Example 11.1.1**

Blood Pressure. Suppose that each of 10 patients is treated with the same amount of two different drugs that can affect blood pressure. To be specific, each patient is first treated with a standard drug $A$, and their change in blood pressure is measured. After the effect of the drug wears off, the patient is treated with an equal amount of a new drug $B$, and their change in blood pressure is measured again. These changes in blood pressure will be called the *reaction* of the patient to each drug. For $i = 1, \ldots, 10$, we shall let $x_i$ denote the reaction, measured in appropriate units, of the $i$th patient to drug $A$, and we shall let $y_i$ denote her reaction to drug $B$. The observed values of the reactions are as given in Table 11.1. The 10 points $(x_i, y_i)$ for $i = 1, \ldots, 10$ are plotted in Fig. 11.1. One purpose of the study is to try to predict a patient's reaction to drug $B$ if their reaction to the standard drug $A$ is already known. ◀

In Example 11.1.1, suppose that we are interested in describing the relationship between the reaction $y$ of a patient to drug $B$ and her reaction $x$ to drug $A$. In order to obtain a simple expression for this relationship, we might wish to fit a straight line to the 10 points plotted in Fig. 11.1. Although these 10 points obviously do not lie exactly on a straight line, we might believe that the deviations from such a line are caused by the fact that the observed change in the blood pressure of each patient is affected not only by the two drugs but also by various other factors. In other words, we might believe that if it were possible to control all of these other factors, the observed points would actually lie on a straight line. We might believe further that if we measured the reactions to the two drugs for a very large number of patients, instead of for just 10 patients, we would then find that the observed points tend to

**Table 11.1** Reactions to two drugs

| $i$ | $x_i$ | $y_i$ |
|---|---|---|
| 1 | 1.9 | 0.7 |
| 2 | 0.8 | −1.0 |
| 3 | 1.1 | −0.2 |
| 4 | 0.1 | −1.2 |
| 5 | −0.1 | −0.1 |
| 6 | 4.4 | 3.4 |
| 7 | 4.6 | 0.0 |
| 8 | 1.6 | 0.8 |
| 9 | 5.5 | 3.7 |
| 10 | 3.4 | 2.0 |

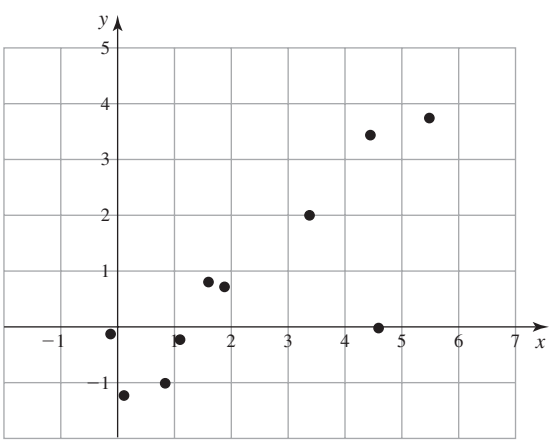


**Figure 11.1** A plot of the observed values in Table 11.1.

cluster along a straight line. Perhaps we might also wish to be able to predict the reaction $y$ of a future patient to the new drug $B$ on the basis of her reaction $x$ to the standard drug $A$. One procedure for making such a prediction would be to fit a straight line to the points in Fig. 11.1, and to use this line for predicting the value of $y$ corresponding to each value of $x$.

It can be seen from Fig. 11.1 that if we did not have to consider the point (4.6, 0.0), which is obtained from the patient for whom $i = 7$ in Table 11.1, then the other nine points lie roughly along a straight line. One arbitrary line that fits reasonably well to these nine points is sketched in Fig. 11.2. However, if we wish to fit a straight line to all 10 points, it is not clear just how much the line in Fig. 11.2 should be adjusted in order to accommodate the anomalous point. We shall now describe a method for fitting such a line.

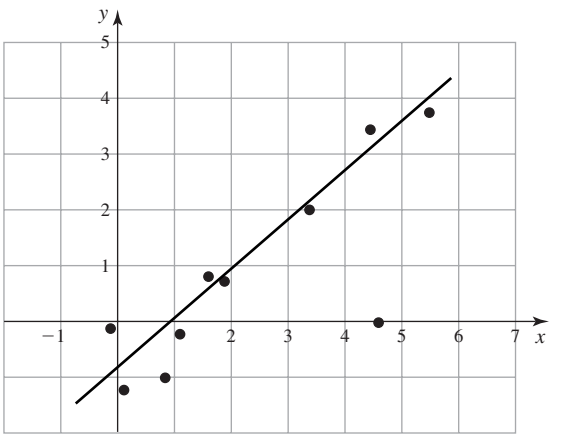


**Figure 11.2** A straight line fitted to nine of the points in Table 11.1.

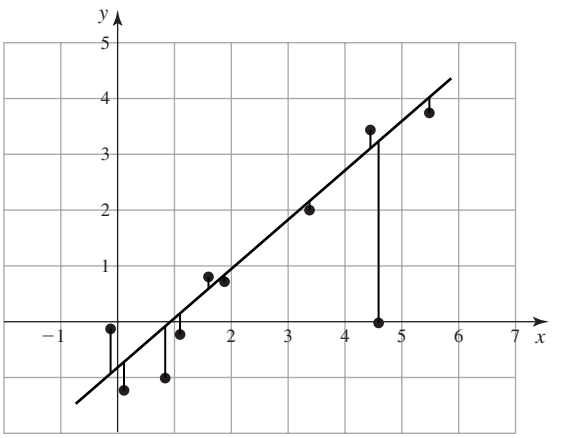


**Figure 11.3** Vertical deviations of the plotted points from a straight line.

## The Least-Squares Line

**Example 11.1.2**

Blood Pressure. In Example 11.1.1, suppose that we are interested in fitting a straight line to the points plotted in Fig. 11.1 in order to obtain a simple mathematical relationship for expressing the reaction $y$ of a patient to the new drug $B$ as a function of her reaction $x$ to the standard drug $A$. In other words, our main objective is to be able to predict closely a patient's reaction $y$ to drug $B$ from her reaction $x$ to drug $A$. We are interested, therefore, in constructing a straight line such that, for each observed reaction $x_i$, the corresponding value of $y$ on the straight line will be as close as possible to the actual observed reaction $y_i$. The vertical deviations of the 10 plotted points from the line drawn in Fig. 11.2 are sketched in Fig. 11.3. ◀

One method of constructing a straight line to fit the observed values is called *the method of least squares*, which chooses the line to minimize the sum of the squares of the vertical deviations of all the points from the line. We shall now study the method of least squares in more detail.

**Theorem 11.1.1** Least Squares. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a set of $n$ points. The straight line that minimzes the sum of the squares of the vertical deviations of all the points from the line has the following slope and intercept:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}, \tag{11.1.1}$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

**Proof** Consider an arbitrary straight line $y = \beta_0 + \beta_1 x$, in which the values of the constants $\beta_0$ and $\beta_1$ are to be determined. When $x = x_i$, the height of this line is $\beta_0 + \beta_1 x_i$. Therefore, the vertical distance between the point $(x_i, y_i)$ and the line is $|y_i - (\beta_0 + \beta_1 x_i)|$. Suppose that the line is to be fitted to $n$ points. The sum of the squares of the vertical distances at the $n$ points is

$$Q = \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2. \tag{11.1.2}$$

We shall minimize $Q$ with respect to $\beta_0$ and $\beta_1$ by taking the partial derivatives and setting them to 0. We have

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \tag{11.1.3}$$

and

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i. \tag{11.1.4}$$

By setting each of these two partial derivatives equal to 0, we obtain the following pair of equations:

$$\begin{aligned} \beta_0 n + \beta_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i, \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i. \end{aligned} \tag{11.1.5}$$

The equations (11.1.5) are called the *normal equations* for $\beta_0$ and $\beta_1$. By considering the second-order derivatives of $Q$, we can show that the values of $\beta_0$ and $\beta_1$ that satisfy the normal equations will be the values for which the sum of squares $Q$ in Eq. (11.1.2) is minimized. Solving (11.1.5) yields the values in (11.1.1). ■

**Definition 11.1.1** Least-Squares Line. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be as defined in (11.1.1). The line defined by the equation $y = \hat{\beta}_0 + \hat{\beta}_1 x$ is called the *least-squares line*.

For the values given in Table 11.1, $n = 10$, and it is found from Eq. (11.1.1) that $\hat{\beta}_0 = -0.786$ and $\hat{\beta}_1 = 0.685$. Hence, the equation of the least-squares line is $y = -0.786 + 0.685x$. This line is sketched in Fig. 11.4.

Virtually all statistical computer software will compute the least-squares regression line. Even some handheld calculators will do the calculation.

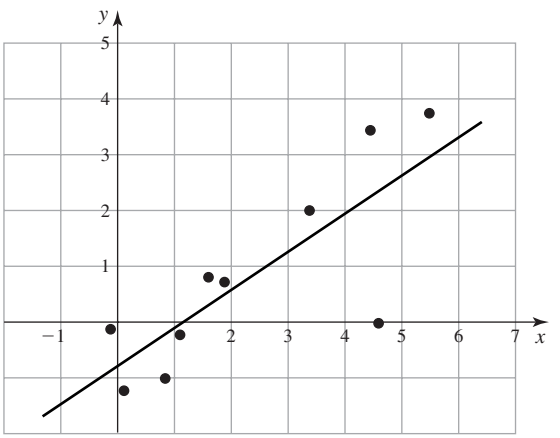


**Figure 11.4** The least-squares straight line.

## Fitting a Polynomial by the Method of Least Squares

Suppose now that instead of simply fitting a straight line to $n$ plotted points, we wish to fit a polynomial of degree $k$ ($k \geq 2$). Such a polynomial will have the following form:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k. \tag{11.1.6}$$

The method of least squares specifies that the constants $\beta_0, \ldots, \beta_k$ should be chosen so that the sum $Q$ of the squares of the vertical deviations of the points from the curve is a minimum. In other words, these constants should be chosen so as to minimize the following expression for $Q$:

$$Q = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i + \cdots + \beta_k x_i^k)]^2. \tag{11.1.7}$$

If we calculate the $k+1$ partial derivatives $\partial Q/\partial\beta_0, \ldots, \partial Q/\partial\beta_k$, and we set each of these derivatives equal to 0, we obtain the following $k+1$ linear equations involving the $k+1$ unknown values $\beta_0, \ldots, \beta_k$:

$$\begin{aligned}
\beta_0 n + \beta_1 \sum_{i=1}^{n} x_i + \cdots + \beta_k \sum_{i=1}^{n} x_i^k &= \sum_{i=1}^{n} y_i, \\
\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 + \cdots + \beta_k \sum_{i=1}^{n} x_i^{k+1} &= \sum_{i=1}^{n} x_i y_i, \\
&\vdots \\
\beta_0 \sum_{i=1}^{n} x_i^k + \beta_1 \sum_{i=1}^{n} x_i^{k+1} + \cdots + \beta_k \sum_{i=1}^{n} x_i^{2k} &= \sum_{i=1}^{n} x_i^k y_i.
\end{aligned} \tag{11.1.8}$$

As before, these equations are called the *normal equations*. If the normal equations have a unique solution, that solution provides the minimum value for $Q$. A necessary and sufficient condition for a unique solution is that the determinant of the $(k+1) \times (k+1)$ matrix formed by the coefficients of $\beta_0, \ldots, \beta_k$ in Eq. (11.1.8) is not zero. We shall now assume that this is the case. If we denote the solution as $(\hat{\beta}_0, \ldots, \hat{\beta}_k)$, then the least-squares polynomial is $y = \hat{\beta}_0 + \hat{\beta}_1 x + \cdots + \hat{\beta}_k x^k$.

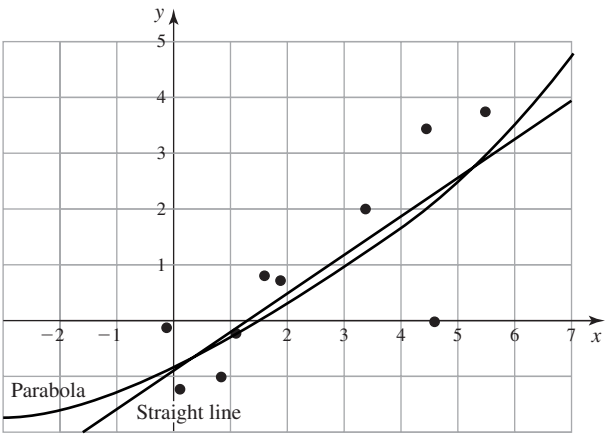


**Figure 11.5** The least-squares parabola.

**Example 11.1.3**

**Fitting a Parabola.** Suppose that we wish to fit a polynomial of the form $y = \beta_0 + \beta_1 x + \beta_2 x^2$ (which represents a parabola) to the 10 points given in Table 11.1. In this example, it is found that the normal equations 11.1.8 are as follows:

$$\begin{aligned} 10\beta_0 + 23.3\beta_1 + 90.37\beta_2 &= 8.1, \\ 23.3\beta_0 + 90.37\beta_1 + 401.0\beta_2 &= 43.59, \\ 90.37\beta_0 + 401.0\beta_1 + 1892.7\beta_2 &= 204.55. \end{aligned} \tag{11.1.9}$$

The unique values of $\beta_0$, $\beta_1$, and $\beta_2$ that satisfy these three equations are $\hat{\beta}_0 = -0.744$, $\hat{\beta}_1 = 0.616$, and $\hat{\beta}_2 = 0.013$. Hence, the least-squares parabola is

$$y = -0.744 + 0.616x + 0.013x^2. \tag{11.1.10}$$

This curve is sketched in Fig. 11.5 together with the least-squares straight line. Because the coefficient of $x^2$ in Eq. (11.1.10) is so small, the least-squares parabola and the least-squares straight line are very close together over the range of values included in Fig. 11.5. ◀

**Example 11.1.4**

**Gasoline Mileage.** Heavenrich and Hellman (1999) report several variables measured on 173 different cars. Among those variables are gasoline mileage (in miles per gallon) and engine horsepower. A plot of miles per gallon versus horsepower is shown in Fig. 11.6 together with a parabola fit by least squares. Even without the curve

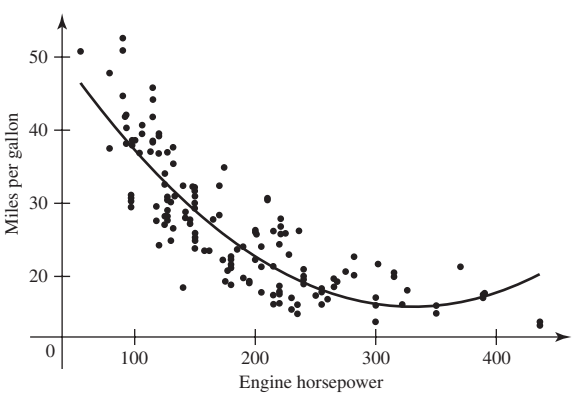


**Figure 11.6** Plot of miles per gallon versus engine horsepower for 173 cars in Example 11.1.4. The least-squares parabola is also drawn in the plot.

drawn in Fig. 11.6, it is clear that a straight line would not provide an adequate fit to the relationship between these two variables. Some sort of curved relationship must be fit. The least-squares parabola curves up for the largest values of horsepower, which is somewhat counterintuitive. Indeed, this might be an example in which it would pay to use some prior information to impose a constraint on the fitted curve. Alternatively, we could replace gasoline mileage by a curved function of miles per gallon and use this curved function as the $y$ variable. ◀

## Fitting a Linear Function of Several Variables

We shall now consider an extension of the example discussed at the beginning of this section, in which we were interested in representing a patient's reaction to a new drug $B$ as a linear function of her reaction to drug $A$. Suppose that we wish to represent a patient's reaction to drug $B$ as a linear function involving not only her reaction to drug $A$ but also some other relevant variables. For example, we may wish to represent the patient's reaction $y$ to drug $B$ as a linear function involving her reaction $x_1$ to drug $A$, her heart rate $x_2$, and blood pressure $x_3$ before she receives any drugs, and other relevant variables $x_4, \ldots, x_k$.

Suppose that for each patient $i$ $(i = 1, \ldots, n)$ we measure her reaction $y_i$ to drug $B$, her reaction $x_{i1}$ to drug $A$, and also her values $x_{i2}, \ldots, x_{ik}$ for the other variables. Suppose also that in order to fit these observed values for the $n$ patients, we wish to consider a linear function having the form

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{11.1.11}$$

In this case, also, the values of $\beta_0, \ldots, \beta_k$ can be determined by the method of least squares. For each given set of observed values $x_{i1}, \ldots, x_{ik}$, we again consider the difference between the observed reaction $y_i$ and the value $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$ of the linear function given in Eq. (11.1.11). As before, it is required to minimize the sum $Q$ of the squares of these differences. Here,

$$Q = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})]^2. \tag{11.1.12}$$

We minimize this the same way that we minimized (11.1.7), namely, by setting the partial derivatives of $Q$ with respect to each $\beta_j$ equal to 0 for $j = 0, \ldots, k$. In this case, the $k + 1$ normal equations have the following form:

$$\begin{aligned}
\beta_0 n + \beta_1 \sum_{i=1}^{n} x_{i1} + \cdots + \beta_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i, \\
\beta_0 \sum_{i=1}^{n} x_{i1} + \beta_1 \sum_{i=1}^{n} x_{i1}^2 + \cdots + \beta_k \sum_{i=1}^{n} x_{i1} x_{ik} &= \sum_{i=1}^{n} x_{i1} y_i, \\
&\vdots \\
\beta_0 \sum_{i=1}^{n} x_{ik} + \beta_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \cdots + \beta_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik} y_i.
\end{aligned} \tag{11.1.13}$$

If the normal equations have a unique solution, we shall denote that solution $(\hat{\beta}_0, \ldots, \hat{\beta}_k)$, and the least-squares linear function will then be $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$. As before, a necessary and sufficient condition for a unique solution is that the determinant of the $(k + 1) \times (k + 1)$ matrix formed by the coefficients of $\beta_0, \ldots, \beta_k$ in Eq. (11.1.13) is not zero.

**Table 11.2** Reactions to two drugs and heart rate

| $i$ | $x_{i1}$ | $x_{i2}$ | $y_i$ |
|---|---|---|---|
| 1 | 1.9 | 66 | 0.7 |
| 2 | 0.8 | 62 | −1.0 |
| 3 | 1.1 | 64 | −0.2 |
| 4 | 0.1 | 61 | −1.2 |
| 5 | −0.1 | 63 | −0.1 |
| 6 | 4.4 | 70 | 3.4 |
| 7 | 4.6 | 68 | 0.0 |
| 8 | 1.6 | 62 | 0.8 |
| 9 | 5.5 | 68 | 3.7 |
| 10 | 3.4 | 66 | 2.0 |

**Example 11.1.5** Fitting a Linear Function of Two Variables. Suppose that we expand Table 11.1 to include the values given in the third column in Table 11.2. Here, for each patient $i$ ($i = 1, \ldots, 10$), $x_{i1}$ denotes her reaction to the standard drug $A$, $x_{i2}$ denotes her heart rate, and $y_i$ denotes her reaction to the new drug $B$. Suppose also that we wish to fit a linear function to these values having the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

In this example, it is found that the normal equations (11.1.13) are

$$\begin{aligned} 10\beta_0 + 23.3\beta_1 + 650\beta_2 &= 8.1, \\ 23.3\beta_0 + 90.37\beta_1 + 1563.6\beta_2 &= 43.59, \\ 650\beta_0 + 1563.6\beta_1 + 42,334\beta_2 &= 563.1. \end{aligned} \tag{11.1.14}$$

The unique values of $\beta_0$, $\beta_1$, and $\beta_2$ that satisfy these three equations are $\hat{\beta}_0 = -11.4527$, $\hat{\beta}_1 = 0.4503$, and $\hat{\beta}_2 = 0.1725$. Hence, the least-squares linear function is

$$y = -11.4527 + 0.4503x_1 + 0.1725x_2. \tag{11.1.15}$$

◀

It should be noted that the problem of fitting a polynomial of degree $k$ involving only one variable, as specified by Eq. (11.1.6), can be regarded as a special case of the problem of fitting a linear function involving several variables, as specified by Eq. (11.1.11). To make Eq. (11.1.11) applicable to the problem of fitting a polynomial having the form given in Eq. (11.1.6), we define the $k$ variables $x_1, \ldots, x_k$ simply as $x_1 = x, x_2 = x^2, \ldots, x_k = x^k$.

A polynomial involving more than one variable can also be represented in the form of Eq. (11.1.11). For example, suppose that the values of four variables $r$, $s$, $t$, and $y$ are observed for several different patients, and we wish to fit to these observed values a function having the following form:

$$y = \beta_0 + \beta_1 r + \beta_2 r^2 + \beta_3 rs + \beta_4 s^2 + \beta_5 t^3 + \beta_6 rst. \tag{11.1.16}$$

We can regard the function in Eq. (11.1.16) as a linear function having the form given in Eq. (11.1.11) with $k = 6$ if we define the six variables $x_1, \ldots, x_6$ as follows: $x_1 = r$, $x_2 = r^2$, $x_3 = rs$, $x_4 = s^2$, $x_5 = t^3$, and $x_6 = rst$.

## Summary

The method of least squares allows the calculation of a predictor for one variable ($y$) based on one or more other variables ($x_1, \ldots, x_k$) of the form $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$. The coefficients $\beta_0, \ldots, \beta_k$ are chosen so that the sum of squared differences between observed values of $y$ and observed values of $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ is as small as possible. Algebraic formulas for the coefficients are given for the case $k = 1$, but most statistical computer software will calculate the coefficients more easily.

## Exercises

**1.** Prove that $\sum_{i=1}^{n}(c_1 x_i + c_2)^2 = c_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(c_1\overline{x} + c_2)^2$.

**2.** Show that the value of $\hat{\beta}_1$ in Eq. (11.1.1) can be rewritten in each of the following three forms:

**a.** $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$

**b.** $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$

**c.** $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$

**3.** Show that the least-squares line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ passes through the point $(\overline{x}, \overline{y})$.

**4.** For $i = 1, \ldots, n$, let $\hat{y}_i = \beta_0 + \beta_1 x_i$. Show that $\hat{\beta}_0$ and $\hat{\beta}_1$, as given by Eq. (11.1.1), are the unique values of $\beta_0$ and $\beta_1$ such that

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i(y_i - \hat{y}_i) = 0.$$

**5.** Fit a straight line to the observed values given in Table 11.1 so that the sum of the squares of the *horizontal* deviations of the points from the line is a minimum. Sketch on the same graph both this line and the least-squares line given in Fig. 11.4.

**6.** Suppose that both the least-squares line and the least-squares parabola were fitted to the same set of points. Explain why the sum of the squares of the deviations of the points from the parabola cannot be larger than the sum of the squares of the deviations of the points from the straight line.

**7.** Suppose that eight specimens of a certain type of alloy were produced at different temperatures, and the durability of each specimen was then observed. The observed values are given in Table 11.3, where $x_i$ denotes the temperature (in coded units) at which specimen $i$ was produced and $y_i$ denotes the durability (in coded units) of that specimen.

**Table 11.3** Data for Exercise 7

| $i$ | $x_i$ | $y_i$ |
|---|---|---|
| 1 | 0.5 | 40 |
| 2 | 1.0 | 41 |
| 3 | 1.5 | 43 |
| 4 | 2.0 | 42 |
| 5 | 2.5 | 44 |
| 6 | 3.0 | 42 |
| 7 | 3.5 | 43 |
| 8 | 4.0 | 42 |

**a.** Fit a straight line of the form $y = \beta_0 + \beta_1 x$ to these values by the method of least squares.

**b.** Fit a parabola of the form $y = \beta_0 + \beta_1 x + \beta_2 x^2$ to these values by the method of least squares.

**c.** Sketch on the same graph the eight data points, the line found in part (a), and the parabola found in part (b).

**8.** Let $(x_i, y_i)$ for $i = 1, \ldots, k + 1$, denote $k + 1$ given points in the $xy$-plane such that no two of these points have the same $x$-coordinate. Show that there is a unique polynomial having the form $y = \beta_0 + \beta_1 x + \cdots + \beta_k x^k$ that passes through these $k + 1$ points.

**9.** The resilience $y$ of a certain type of plastic is to be represented as a linear function of both the temperature $x_1$ at which the plastic is baked and the number of minutes $x_2$ for which it is baked. Suppose that 10 pieces of plastic are prepared by using different values of $x_1$ and $x_2$, and the observed values in appropriate units are as given in Table 11.4. Fit a function having the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ to these observed values by the method of least squares.

**10.** Consider again the observed values presented in Table 11.4. Fit a function having the form $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2$ to these values by the method of least squares.

**Table 11.4** Data for Exercise 9

| $i$ | $x_{i1}$ | $x_{i2}$ | $y_i$ | $i$ | $x_{i1}$ | $x_{i2}$ | $y_i$ |
|---|---|---|---|---|---|---|---|
| 1 | 100 | 1 | 113 | 6 | 120 | 2 | 144 |
| 2 | 100 | 2 | 118 | 7 | 120 | 3 | 138 |
| 3 | 110 | 1 | 127 | 8 | 130 | 1 | 146 |
| 4 | 110 | 2 | 132 | 9 | 130 | 2 | 156 |
| 5 | 120 | 1 | 136 | 10 | 130 | 3 | 149 |

**11.** Consider again the observed values presented in Table 11.4, and consider also the two functions that were fitted to these values in Exercises 9 and 10. Which of these two functions fits the observed values better?

## 11.2 Regression

*In Sec. 11.1, we introduced the method of least squares. This method computes coefficients for a linear function to predict one variable y based on other variables $x_1, \ldots, x_k$. In this section, we assume that the y values are observed values of a collection of random variables. In this case, there is a statistical model in which the method of least squares turns out to produce the maximum likelihood estimates of the parameters of the model.*

### Regression Functions

**Example 11.2.1**

Pressure and the Boiling Point of Water. Forbes (1857) reports the results from experiments that were trying to obtain a method for estimating altitude. A formula is available for altitude in terms of barometric pressure, but it was difficult to carry a barometer to high altitudes in Forbes' day. However, it might be easy for travelers to carry a thermometer and measure the boiling point of water. Table 11.5 contains the measured barometric pressures and boiling points of water from 17 experiments. We can use the method of least squares to fit a linear relationship between boiling point and pressure. Let $y_i$ be the pressure for one of Forbes' observations, and let $x_i$ be the corresponding boiling point for $i = 1, \ldots, 17$. Using the data in Table 11.5, we can compute the least-squares line. The intercept and slope are, respectively, $\hat{\beta}_0 = -81.049$ and $\hat{\beta}_1 = 0.5228$. Of course, we do not expect that the line $y = -81.049 + 0.5228x$ precisely gives the relationship between boiling point $x$ and pressure $y$. If we learn the boiling point $x$ of water and want to compute the conditional distribution of the unknown pressure $Y$, is there a statistical model that allows us to say what the (conditional) distribution of pressure is given that the boiling point is $x$? ◀

In this section, we shall describe a statistical model for problems such as the one in Example 11.2.1. Fitting this statistical model will make use of the method of least squares. We shall study problems in which we are interested in learning about the conditional distribution of some random variable $Y$ for given values of some other variables $X_1, \ldots, X_k$. The variables $X_1, \ldots, X_k$ may be random variables whose values are to be observed in an experiment along with the values of $Y$, or they may be *control variables* whose values are to be chosen by the experimenter. In general, some

**Table 11.5** Boiling point of water in degrees Fahrenheit and atmospheric pressure in inches of mercury from Forbes' experiments. These data are taken from Weisberg (1985, p. 3).

| Boiling Point | Pressure |
|---|---|
| 194.5 | 20.79 |
| 194.3 | 20.79 |
| 197.9 | 22.40 |
| 198.4 | 22.67 |
| 199.4 | 23.15 |
| 199.9 | 23.35 |
| 200.9 | 23.89 |
| 201.1 | 23.99 |
| 201.4 | 24.02 |
| 201.3 | 24.01 |
| 203.6 | 25.14 |
| 204.6 | 26.57 |
| 209.5 | 28.49 |
| 208.6 | 27.76 |
| 210.7 | 29.04 |
| 211.9 | 29.88 |
| 212.2 | 30.06 |

of these variables might be random variables, and some might be control variables. In any case, we can study the conditional distribution of $Y$ given $X_1, \ldots, X_k$. We begin with some terminology.

**Definition 11.2.1** Response/Predictor/Regression. The variables $X_1, \ldots, X_k$ are called *predictors*, and the random variable $Y$ is called the *response*. The conditional expectation of $Y$ for given values $x_1, \ldots, x_k$ of $X_1, \ldots, X_k$ is called the *regression function of $Y$ on $X_1, \ldots, X_k$*, or simply the *regression of $Y$ on $X_1, \ldots, X_k$*.

The regression of $Y$ on $X_1, \ldots, X_k$ is a function of the values $x_1, \ldots, x_k$ of $X_1, \ldots, X_k$. In symbols, this function is $E(Y|x_1, \ldots, x_k)$.

In this chapter, we shall assume that the regression function $E(Y|x_1, \ldots, x_k)$ is a linear function having the following form:

$$E(Y|x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{11.2.1}$$

The coefficients $\beta_0, \ldots, \beta_k$ in Eq. (11.2.1) are called *regression coefficients*. We shall suppose that these regression coefficients are unknown. Therefore, they are to be regarded as parameters whose values are to be estimated. We shall suppose also that $n$ vectors of observations are obtained. For $i = 1, \ldots, n$, we shall assume that the $i$th vector $(x_{i1}, \ldots, x_{ik}, y_i)$ consists of a set of controlled or observed values of $X_1, \ldots, X_k$ and the corresponding observed value of $Y$.

One set of estimators of the regression coefficients $\beta_0, \ldots, \beta_k$ that can be calculated from these observations is the set of values $\hat{\beta}_0, \ldots, \hat{\beta}_k$ that are obtained by the method of least squares, as described in Sec. 11.1. These estimators are called the *least-squares estimators* of $\beta_0, \ldots, \beta_k$. We shall now specify some further assumptions about the conditional distribution of $Y$ given $X_1, \ldots, X_k$ in order to be able to determine in greater detail the properties of these least-squares estimators.

## Simple Linear Regression

We shall consider first a problem in which we wish to study the regression of $Y$ on just a single variable $X$. We shall assume that for each value $X = x$, the random variable $Y$ can be represented in the form $Y = \beta_0 + \beta_1 x + \varepsilon$, where $\varepsilon$ is a random variable that has the normal distribution with mean 0 and variance $\sigma^2$. It follows from this assumption that the conditional distribution of $Y$ given $X = x$ is the normal distribution with mean $\beta_0 + \beta_1 x$ and variance $\sigma^2$.

A problem of this type is called a problem of *simple linear regression*. Here the term *simple* refers to the fact that we are considering the regression of $Y$ on just a single variable $X$, rather than on more than one variable; the term *linear* refers to the fact that the regression function $E(Y|x) = \beta_0 + \beta_1 x$ is a linear function of the parameters $\beta_0$ and $\beta_1$. For example, a problem in which $E(Y|x)$ is a polynomial, like the right side of Eq. (11.1.6), would also be a linear regression problem, but not simple.

Throughout this section (and the next two sections), we shall consider the problem in which we shall observe $n$ pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$. We shall make the following five assumptions. Each of these assumptions has a natural generalization to the case in which there is more than one predictor, but we shall postpone discussion of that case until Sec. 11.5.

**Assumption 11.2.1** Predictor is known. Either the values $x_1, \ldots, x_n$ are known ahead of time or they are the observed values of random variables $X_1, \ldots, X_n$ on whose values we condition before computing the joint distribution of $(Y_1, \ldots, Y_n)$.

**Assumption 11.2.2** Normality. For $i = 1, \ldots, n$, the conditional distribution of $Y_i$ given the values $x_1, \ldots, x_n$ is a normal distribution.

**Assumption 11.2.3** Linear Mean. There are parameters $\beta_0$ and $\beta_1$ such that the conditional mean of $Y_i$ given the values $x_1, \ldots, x_n$ has the form $\beta_0 + \beta_1 x_i$ for $i = 1, \ldots, n$.

**Assumption 11.2.4** Common Variance. There is a parameter $\sigma^2$ such that the conditional variance of $Y_i$ given the values $x_1, \ldots, x_n$ is $\sigma^2$ for $i = 1, \ldots, n$. This assumption is often called *homoscedasticity*. Random variables with different variances are called *heteroscedastic*.

**Assumption 11.2.5** Independence. The random variables $Y_1, \ldots, Y_n$ are independent given the observed $x_1, \ldots, x_n$.

A brief word is in order about Assumption 11.2.1. In Example 11.1.1, we saw that the reaction $x_i$ of patient $i$ to standard drug $A$ is observed as part of the experiment along with the reaction $y_i$ to drug $B$. Hence, the predictors are not known in advance. In this case, all probability statements that we make in this example are conditional on $(x_1, \ldots, x_n)$. In other examples, one might be trying to predict an economic variable using the year in which it was measured. In such cases, such as Example 11.5.1, which

we will see later, the values of at least some of the predictors are truely known in advance.

Assumptions 11.2.1–11.2.5 specify the conditional joint distribution of $Y_1, \ldots, Y_n$ given the vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ and the parameters $\beta_0$, $\beta_1$, and $\sigma^2$. In particular, the conditional joint p.d.f. of $Y_1, \ldots, Y_n$ is

$$f_n(\boldsymbol{y}|\boldsymbol{x}, \beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]. \tag{11.2.2}$$

We can now find the M.L.E.'s of $\beta_0$, $\beta_1$, and $\sigma^2$.

**Theorem 11.2.1** **Simple Linear Regression M.L.E.'s.** Assume Assumptions 11.2.1–11.2.5. The M.L.E.'s of $\beta_0$ and $\beta_1$ are the least-squares estimates, and the M.L.E. of $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \tag{11.2.3}$$

**Proof** For each observed vector $\boldsymbol{y} = (y_1, \ldots, y_n)$, the p.d.f. (11.2.2) will be the likelihood function of the parameters $\beta_0$, $\beta_1$, and $\sigma^2$. In Eq. (11.2.2), $\beta_0$ and $\beta_1$ appear only in the sum of squares

$$Q = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2,$$

which in turn appears in the exponent multiplied by $-1/[2\sigma^2]$. Regardless of the value of $\sigma^2$, the exponent is maximized over $\beta_0$ and $\beta_1$ by minimizing $Q$. It follows that the M.L.E.'s can be found in sequence by first minimizing $Q$ over $\beta_0$ and $\beta_1$, then inserting the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that provide the minimum of $Q$, and finally minimizing the result over $\sigma^2$. The reader will note that $Q$ is the same as the sum of squares in Eq. (11.1.2), which is minimized by the method of least squares. Thus, the M.L.E.'s of the regression coefficients $\beta_0$ and $\beta_1$ are precisely the same as the least-squares estimates. The exact form of these estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ was given in Eq. (11.1.1).

To find the M.L.E. of $\sigma^2$, perform the the second and third steps described in the preceding paragraph, namely, first replace $\beta_0$ and $\beta_1$ in Eq. (11.2.2) by their M.L.E.'s $\hat{\beta}_0$ and $\hat{\beta}_1$, and then maximize the resulting expression with respect to $\sigma^2$. The details are left to Exercise 1 at the end of this section, and the result is (11.2.3). ■

## The Distribution of the Least-Squares Estimators

We shall now present the joint distribution of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ when they are regarded as functions of the random variables $Y_1, \ldots, Y_n$ for given values of $x_1, \ldots, x_n$. Specifically, the estimators are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x},$$

where $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

It is convenient, both for this section and the next, to introduce the symbol

$$s_x = \left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{1/2}. \tag{11.2.4}$$

**Theorem 11.2.2** Distributions of Least-Squares Estimators. Under Assumptions 11.2.1–11.2.5, the distribution of $\hat{\beta}_1$ is the normal distribution with mean $\beta_1$ and variance $\sigma^2/s_x^2$. The distribution of $\hat{\beta}_0$ is the normal distribution with mean $\beta_0$ and variance

$$\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{s_x^2}\right). \tag{11.2.5}$$

Finally, the covariance of $\hat{\beta}_1$ and $\hat{\beta}_0$ is

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\overline{x}\sigma^2}{s_x^2}. \tag{11.2.6}$$

(All of the distributional statements in this theorem are conditional on $X_i = x_i$ for $i = 1, \ldots, n$ if $X_1, \ldots, X_n$ are random variables.)

**Proof** To determine the distribution of $\hat{\beta}_1$, it is convenient to write $\hat{\beta}_1$ as follows (see Exercise 2 at the end of Sec. 11.1):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{s_x^2}. \tag{11.2.7}$$

It can be seen from Eq. (11.2.7) that $\hat{\beta}_1$ is a linear function of $Y_1, \ldots, Y_n$. Because the random variables $Y_1, \ldots, Y_n$ are independent and each has a normal distribution, it follows that $\hat{\beta}_1$ will also have a normal distribution. Furthermore, the mean of this distribution will be

$$E(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \overline{x})E(Y_i)}{s_x^2}.$$

Because $E(Y_i) = \beta_0 + \beta_1 x_i$ for $i = 1, \ldots, n$, it can now be found (see Exercise 2 at the end of this section) that

$$E(\hat{\beta}_1) = \beta_1. \tag{11.2.8}$$

Furthermore, because the random variables $Y_1, \ldots, Y_n$ are independent and each has variance $\sigma^2$, it follows from Eq. (11.2.7) that

$$\text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \overline{x})^2 \text{Var}(Y_i)}{s_x^4} = \frac{\sigma^2}{s_x^2}. \tag{11.2.9}$$

Next, consider the distribution of $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{x}$. Because both $\overline{Y}$ and $\hat{\beta}_1$ are linear functions of $Y_1, \ldots, Y_n$, it follows that $\hat{\beta}_0$ is also a linear function of $Y_1, \ldots, Y_n$. Hence, $\hat{\beta}_0$ will have a normal distribution. The mean of $\hat{\beta}_0$ can be determined from the relation $E(\hat{\beta}_0) = E(\overline{Y}) - \overline{x}E(\hat{\beta}_1)$. It can be shown (see Exercise 3) that $E(\hat{\beta}_0) = \beta_0$. Furthermore, it can be shown (see Exercise 4) that $\text{Var}(\hat{\beta}_0)$ is given by (11.2.5). Finally, it can be shown (see Exercise 5) that the value of the covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$ is given by (11.2.6). ■

A simple corollary to Theorem 11.2.2 is that $\hat{\beta}_0$ and $\hat{\beta}_1$ are, respectively, unbiased estimators of the corresponding parameters $\beta_0$ and $\beta_1$.

To complete the description of the joint distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$, it will be shown in Sec. 11.3 that this joint distribution is the bivariate normal distribution for which the means, variances, and covariance are as stated in Theorem 11.2.2.

**Example 11.2.2** Pressure and the Boiling Point of Water. In Example 11.2.1, we found the least-squares line for predicting pressure from boiling point of water. Suppose that we use the linear regression model just described as a model for the data in this experiment. That is, let $Y_i$ be the pressure for one of Forbes' observations, and let $x_i$ be the corresponding boiling point for $i = 1, \ldots, 17$. We model the $Y_i$ as being independent with means $\beta_0 + \beta_1 x_i$ and variance $\sigma^2$. The average temperature is $\overline{x} = 202.95$ and $s_x^2 = 530.78$ with $n = 17$. From these values, we can now compute the variances and covariances of the least-squares estimators using the formulas derived in this section. For example,

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{530.78} = 0.00188\sigma^2,$$

$$\text{Var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{17} + \frac{202.95^2}{530.78}\right) = 77.66\sigma^2,$$

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{202.95\sigma^2}{530.78} = 0.382\sigma^2.$$

It is easy to see that we expect to get a much more precise estimate of $\beta_1$ than of $\beta_0$. ◀

The statement at the end of Example 11.2.2 about getting more precise estimates of $\beta_1$ than of $\beta_0$ is a bit deceptive. We must multiply $\beta_1$ by a number on the order of 200 before it is on the same scale as $\beta_0$. Hence, it might make more sense to compare the variance of $200\hat{\beta}_1$ to the variance of $\hat{\beta}_0$. In general, we can find the variance of any linear combination of the least-squares estimators.

**Example 11.2.3** The Variance of a Linear Combination. Very often, we need to compute the variance of a linear combination of the least-squares estimators. One example is prediction, as discussed later in this section. Suppose that we wish to compute the variance of $T = c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + c_*$. The variance of $T$ can be found by substituting the values of $\text{Var}(\hat{\beta}_0)$, $\text{Var}(\hat{\beta}_1)$, and $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$ given in Eqs. (11.2.5), (11.2.9), and (11.2.6) in the following relation:

$$\text{Var}(T) = c_0^2\,\text{Var}(\hat{\beta}_0) + c_1^2\,\text{Var}(\hat{\beta}_1) + 2c_0c_1\,\text{Cov}(\hat{\beta}_0, \hat{\beta}_1).$$

When these substitutions have been made, the result can be written in the following form:

$$\text{Var}(T) = \sigma^2\left(\frac{c_0^2}{n} + \frac{(c_0\overline{x} - c_1)^2}{s_x^2}\right). \tag{11.2.10}$$

For the specific case of Example 11.2.2, we have $c_0 = 0$ and $c_1 = 200$, so the variance of $200\hat{\beta}_1$ is $200^2\sigma^2/s_x^2 = 75.36\sigma^2$. This is pretty close to the variance of $\hat{\beta}_0$, namely, $77.66\sigma^2$. ◀

## Prediction

**Example 11.2.4** Predicting Pressure from the Boiling Point of Water. In Example 11.2.1, Forbes was trying to find a way to use the boiling point of water to estimate the barometric pressure. Suppose that a traveler measures the boiling point of water to be 201.5 degrees. What estimate of barometric pressure should they give and how much uncertainty is there about this estimate? ◀

Suppose that $n$ pairs of observations $(x_1, Y_1), \ldots, (x_n, Y_n)$ are to be obtained in a problem of simple linear regression, and on the basis of these $n$ pairs, it is necessary to predict the value of an independent observation $Y$ that will be obtained when a certain specified value $x$ is assigned to the control variable. Since the observation $Y$ will have the normal distribution with mean $\beta_0 + \beta_1 x$ and variance $\sigma^2$, it is natural to use the value $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ as the predicted value of $Y$. We shall now determine the M.S.E. $E[(\hat{Y} - Y)^2]$ of this prediction, where both $\hat{Y}$ and $Y$ are random variables.

**Theorem 11.2.3** M.S.E. of Prediction. In the prediction problem just described,

$$E[(\hat{Y} - Y)^2] = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{s_x^2} \right]. \tag{11.2.11}$$

**Proof** In this problem, $E(\hat{Y}) = E(Y) = \beta_0 + \beta_1 x$. Thus, if we let $\mu = \beta_0 + \beta_1 x$, then

$$\begin{aligned} E[(\hat{Y} - Y)^2] &= E\{[(\hat{Y} - \mu) - (Y - \mu)]^2\} \\ &= \mathrm{Var}(\hat{Y}) + \mathrm{Var}(Y) - 2\,\mathrm{Cov}(\hat{Y}, Y). \end{aligned} \tag{11.2.12}$$

However, the random variables $\hat{Y}$ and $Y$ are independent, because $\hat{Y}$ is a function of the first $n$ pairs of observations and $Y$ is an independent observation. Therefore, $\mathrm{Cov}(\hat{Y}, Y) = 0$, and it follows that

$$E[(\hat{Y} - Y)^2] = \mathrm{Var}(\hat{Y}) + \mathrm{Var}(Y). \tag{11.2.13}$$

Finally, because $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$, the value of $\mathrm{Var}(\hat{Y})$ is given by Eq. (11.2.10) with $c_0 = 1$ and $c_1 = x$. Also $\mathrm{Var}(Y) = \sigma^2$. Substituting these into Eq. (11.2.13) gives (11.2.11). ■

**Example 11.2.5** Predicting Pressure from the Boiling Point of Water. In Example 11.2.4, we wanted to predict barometric pressure when the boiling point of water is 201.5 degrees. The least-squares line is $y = -81.049 + 0.5228x$, and $\hat{\sigma}^2 = 0.0478$. Fig. 11.7 shows the data plotted together with the least-squares regression line and the location of the point on the line that has $x = 201.5$. The M.S.E. of the prediction of pressure $Y$ is obtained from Eq. (11.2.11):

$$E[(\hat{Y} - Y)^2] = \sigma^2 \left[ 1 + \frac{1}{17} + \frac{(201.5 - 202.95)^2}{530.78} \right] = 1.0628\sigma^2,$$

and the observed value of the prediction is $\hat{Y} = -81.06 + 0.5229 \times 201.5 = 24.30$. The calculation of $\hat{Y}$ is illustrated in Fig. 11.7. The M.S.E. $1.0628\sigma^2$ can be interpreted as follows: If we knew the values of $\beta_0$ and $\beta_1$ and tried to predict $Y$, the M.S.E. would be $\mathrm{Var}(Y) = \sigma^2$. Having to estimate $\beta_0$ and $\beta_1$ only costs us an additional $0.0628\sigma^2$ in M.S.E. ◀

**Note: M.S.E. of Prediction Increases as $x$ Moves Away from Observed Data.** The M.S.E. in Eq. (11.2.11) increases as $x$ moves away from $\overline{x}$, and it is smallest when $x = \overline{x}$. This indicates that it is harder to predict $Y$ when $x$ is not near the center of the observed values $x_1, \ldots, x_n$. Indeed, if $x$ is larger than the largest observed $x_i$ or smaller than the smallest one, it is quite difficult to predict $Y$ with much precision. Such predictions outside the range of the observed data are called *extrapolations*.

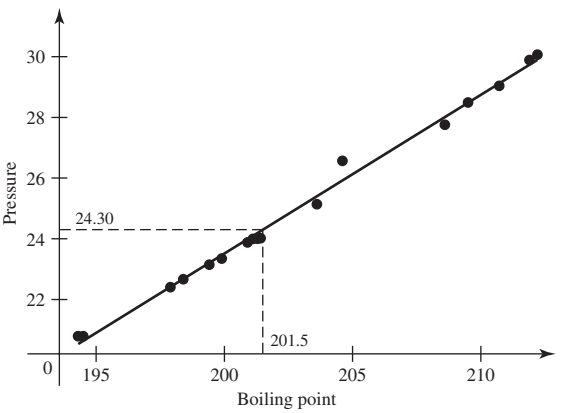


**Figure 11.7** Plot of pressure versus boiling point with regression line for Example 11.2.5. Dotted line illustrates prediction of pressure when boiling point is 201.5.

## Design of the Experiment

Consider a problem of simple linear regression in which the variable $X$ is a control variable whose values $x_1, \ldots, x_n$ can be chosen by the experimenter. We shall discuss methods for choosing these values so as to obtain good estimators of the regression coefficients $\beta_0$ and $\beta_1$.

Suppose first that the values $x_1, \ldots, x_n$ are to be chosen so as to minimize the M.S.E. of the least-squares estimator $\hat{\beta}_0$. Since $\hat{\beta}_0$ is an unbiased estimator of $\beta_0$, the M.S.E. of $\hat{\beta}_0$ is equal to $\text{Var}(\hat{\beta}_0)$, as given in Eq. (11.2.5). It follows from Eq. (11.2.5) that $\text{Var}(\hat{\beta}_0) \geq \sigma^2/n$ for all values $x_1, \ldots, x_n$, and there will be equality in this relation if and only if $\overline{x} = 0$. Hence, $\text{Var}(\hat{\beta}_0)$ will attain its minimum value $\sigma^2/n$ whenever $\overline{x} = 0$. Of course, this will be impossible in any application in which $X$ is constrained to be positive.

Suppose next that the values $x_1, \ldots, x_n$ are to be chosen so as to minimize the M.S.E. of the estimator $\hat{\beta}_1$. Again, the M.S.E. of $\hat{\beta}_1$ will be equal to $\text{Var}(\hat{\beta}_1)$, as given in Eq. (11.2.9). It can be seen from Eq. (11.2.9) that $\text{Var}(\hat{\beta}_1)$ will be minimized by choosing the values $x_1, \ldots, x_n$ so that the value of $s_x^2$ is maximized. If the values $x_1, \ldots, x_n$ must be chosen from some bounded interval $(a, b)$ of the real line, and if $n$ is an even integer, then the value of $s_x^2$ will be maximized by choosing $x_i = a$ for exactly $n/2$ values and choosing $x_i = b$ for the other $n/2$ values. If $n$ is an odd integer, all the values should again be chosen at the endpoints $a$ and $b$, but one endpoint must now receive one more observation than the other endpoint.

It follows from this discussion that if the experiment is to be designed so as to minimize both the M.S.E. of $\hat{\beta}_0$ and the M.S.E. of $\hat{\beta}_1$, then the values $x_1, \ldots, x_n$ should be chosen so that exactly, or approximately, $n/2$ values are equal to some number $c$ that is as large as is feasible in the given experiment, and the remaining values are equal to $-c$. In this way, the value of $\overline{x}$ will be exactly, or approximately, equal to 0, and the value of $s_x^2$ will be as large as possible.

Finally, suppose that the linear combination $\theta = c_0\beta_0 + c_1\beta_1 + c_*$ is to be estimated, where $c_0 \neq 0$, and that the experiment is to be designed so as to minimize the M.S.E. of $\hat{\theta}$, that is, to minimize $\text{Var}(\hat{\theta})$. For example, if $Y$ is a future observation with corresponding predictor $x$, then we could set $c_0 = 1$, $c_2 = x$, and $c_* = 0$ in order to make $\theta = E(Y|x)$. In Example 11.2.3, we computed $\text{Var}(T)$, where $T = \hat{\theta}$, as the sum of two nonnegative terms in Eq. (11.2.10). The second term is the only one that

depends on the values of $x_1, \ldots, x_n$, and it equals 0 (its smallest possible value) if and only if $\overline{x} = c_1/c_0$. In this case, $\text{Var}(\hat{\theta})$ will attain its minimum value $c_0^2\sigma^2/n$.

In practice, an experienced statistician would not usually choose all the values $x_1, \ldots, x_n$ at a single point or at just the two endpoints of the interval $(a, b)$, as the optimal designs that we have just derived would dictate. The reason is that when all $n$ observations are taken at just one or two values of $X$, the experiment provides no possibility of checking the assumption that the regression of $Y$ on $X$ is a linear function. In order to check this assumption without unduly increasing the M.S.E. of the least-squares estimators, many of the values $x_1, \ldots, x_n$ should be chosen at the endpoints $a$ and $b$, but at least some of the values should be chosen at a few interior points of the interval. Linearity can then be checked by visual inspection of the plotted points and the fitting of a polynomial of degree two or higher.

## Summary

We considered the following statistical model. The values $x_1, \ldots, x_n$ are assumed known. The random variables $Y_1, \ldots, Y_n$ are independent with $Y_i$ having the normal distribution with mean $\beta_0 + \beta_1 x_i$ and variance $\sigma^2$. Here, $\beta_0$, $\beta_1$, and $\sigma^2$ are unknown parameters. These are the assumptions of the simple linear regression model. Under this model, the joint distribution of the least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ is a bivariate normal distribution with $\hat{\beta}_i$ having mean $\beta_i$ for $i = 1, 2$. The variances are given in Eqs. (11.2.5) and (11.2.9). The covariance is given in Eq. (11.2.6). If we consider predicting a future $Y$ value with corresponding predictor $x$, we might use the prediction $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. In this case, $Y - \hat{Y}$ has the normal distribution with mean 0 and variance given by Eq. (11.2.11).

## Exercises

**1.** Show that the M.L.E. of $\sigma^2$ is given by Eq. (11.2.3).

**2.** Show that $E(\hat{\beta}_1) = \beta_1$.

**3.** Show that $E(\hat{\beta}_0) = \beta_0$.

**4.** Show that $\text{Var}(\hat{\beta}_0)$ is as given in Eq. (11.2.5).

**5.** Show that $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$ is as given in Eq. (11.2.6). *Hint:* Use the result in Exercise 8 in Sec. 4.6.

**6.** Show that in a problem of simple linear regression, the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ will be independent if $\overline{x} = 0$.

**7.** Consider a problem of simple linear regression in which a patient's reaction $Y$ to a new drug $B$ is to be related to his reaction $X$ to a standard drug $A$. Suppose that the 10 pairs of observed values given in Table 11.1 are obtained.

**a.** Determine the values of the M.L.E.'s $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$.

**b.** Determine the values of $\text{Var}(\hat{\beta}_0)$ and $\text{Var}(\hat{\beta}_1)$.

**c.** Determine the value of the correlation of $\hat{\beta}_0$ and $\hat{\beta}_1$.

**8.** Consider again the conditions of Exercise 7, and suppose that it is desired to estimate the value of $\theta = 3\beta_0 - 2\beta_1 + 5$. Determine an unbiased estimator of $\theta$ and find its M.S.E.

**9.** Consider again the conditions of Exercise 7, and let $\theta = 3\beta_0 + c_1\beta_1$, where $c_1$ is a constant. Determine an unbiased estimator $\hat{\theta}$ of $\theta$. For what value of $c_1$ will the M.S.E. of $\hat{\theta}$ be smallest?

**10.** Consider again the conditions of Exercise 7. If a particular patient's reaction to drug $A$ has the value $x = 2$, what is the predicted value of his reaction to drug $B$, and what is the M.S.E. of this prediction?

**11.** Consider again the conditions of Exercise 7. For what value $x$ of a patient's reaction to drug $A$ can his reaction to drug $B$ be predicted with the smallest M.S.E.?

**12.** Consider a problem of simple linear regression in which the durability $Y$ of a certain type of alloy is to be related to the temperature $X$ at which it was produced. Suppose that the eight pairs of observed values given in Table 11.3 are obtained. Determine the values of the M.L.E.'s $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$, and also the values of $\text{Var}(\hat{\beta}_0)$ and $\text{Var}(\hat{\beta}_1)$.

**13.** For the conditions of Exercise 12, determine the value of the correlation of $\hat{\beta}_0$ and $\hat{\beta}_1$.

**14.** Consider again the conditions of Exercise 12, and suppose that it is desired to estimate the value of $\theta = 5 - 4\beta_0 + \beta_1$. Find an unbiased estimator $\hat{\theta}$ of $\theta$. Determine the value of $\hat{\theta}$ and the M.S.E. of $\hat{\theta}$.

**15.** Consider again the conditions of Exercise 12, and let $\theta = c_1\beta_1 - \beta_0$, where $c_1$ is a constant. Determine an unbiased estimator $\hat{\theta}$ of $\theta$. For what value of $c_1$ will the M.S.E. of $\hat{\theta}$ be smallest?

**16.** Consider again the conditions of Exercise 12. If a specimen of the alloy is to be produced at the temperature $x = 3.25$, what is the predicted value of the durability of the specimen, and what is the M.S.E. of this prediction?

**17.** Consider again the conditions of Exercise 12. For what value of the temperature $x$ can the durability of a specimen of the alloy be predicted with the smallest M.S.E.?

**18.** Moore and McCabe (1999, p. 174) report prices paid for several species of seafood in 1970 and 1980. These values are in Table 11.6. If we were interested in trying to predict 1980 seafood prices from 1970 prices, a linear regression model might be used.

**a.** Find the least-squares regression coefficients for predicting 1980 prices from 1970 prices.

**b.** If an additional species sold for 21.4 in 1970, what would you predict for the 1980 selling price?

**c.** What is the M.S.E. for predicting the 1980 price of a species that sold for 21.4 in 1970?

**Table 11.6** Fish prices in 1970 and 1980 for Exercise 18

| 1970 | 1980 | 1970 | 1980 |
|---|---|---|---|
| 13.1 | 27.3 | 26.7 | 80.1 |
| 15.3 | 42.4 | 47.5 | 150.7 |
| 25.8 | 38.7 | 6.6 | 20.3 |
| 1.8 | 4.5 | 94.7 | 189.7 |
| 4.9 | 23 | 61.1 | 131.3 |
| 55.4 | 166.3 | 135.6 | 404.2 |
| 39.3 | 109.7 | 47.6 | 149 |

**19.** In the 1880s, Francis Galton studied the inheritance of physical characteristics. Galton found that the sons of tall men tended to be taller than average, but shorter than their fathers. Similarly, sons of short men tended to be shorter than average, but taller than their fathers. Thus, the average heights of the sons were closer to the mean height of the population, regardless of whether the fathers were taller or shorter than average. From these observations, one might conclude that the variability of height decreases over successive generations, both tall persons and short persons tend to be eliminated, and the population "regresses" toward some average height. This conclusion is an example of the *regression fallacy*. In this problem you will prove that the regression fallacy arises in the bivariate normal distribution even when both coordinates have the same variance. In particular, assume that the vector $(X_1, X_2)$ has the bivariate normal distribution with common mean $\mu$, common variance $\sigma^2$, and positive correlation $\rho < 1$. Prove that $E(X_2|x_1)$ is closer to $\mu$ than $x_1$ is to $\mu$ for every value $x_1$. (This occurs despite the fact that $X_1$ and $X_2$ have the same mean and the same variance.)