a. Find the value of y so that $r = 1$.
b. Find the value of y so that $r = 0$.
c. Find the value of y so that $r = 0.5$.
d. Find the value of y so that $r = -0.5$.
e. Give a geometric argument to show that there is no value of y for which $r = -1$.

7.2 The Least-Squares Line

When two variables have a linear relationship, the scatterplot tends to be clustered around a line known as the least-squares line (see Figures 7.1 and 7.2 in Section 7.1). In this section we will learn how to compute the least-squares line and how it can be used to draw conclusions from data.

We begin by describing a hypothetical experiment. Springs are used in applications for their ability to extend (stretch) under load. The stiffness of a spring is measured by the "spring constant," which is the length that the spring will be extended by one unit of force or load.[1] To make sure that a given spring functions appropriately, it is necessary to estimate its spring constant with good accuracy and precision.

In our hypothetical experiment, a spring is hung vertically with the top end fixed, and weights are hung one at a time from the other end. After each weight is hung, the length of the spring is measured. Let $x_1, \ldots, x_n$ represent the weights, and let l_i represent the length of the spring under the load x_i. Hooke's law states that

$$l_i = \beta_0 + \beta_1 x_i \tag{7.8}$$

where β_0 is the length of the spring when unloaded and β_1 is the spring constant.

Let y_i be the *measured* length of the spring under load x_i. Because of measurement error, y_i will differ from the true length l_i. We write

$$y_i = l_i + \varepsilon_i \tag{7.9}$$

where ε_i is the error in the ith measurement. Combining (7.8) and (7.9), we obtain

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{7.10}$$

In Equation (7.10) y_i is called the **dependent variable**, x_i is called the **independent variable**, β_0 and β_1 are the **regression coefficients**, and ε_i is called the **error**. Equation (7.10) is called a **linear model**.

Table 7.1 (page 524) presents the results of the hypothetical experiment, and Figure 7.9 (page 524) presents the scatterplot of y versus x. We wish to use these data to estimate the spring constant β_1 and the unloaded length β_0. If there were no measurement error, the points would lie on a straight line with slope β_1 and intercept β_0, and these

[1] The more traditional definition of the spring constant is the reciprocal of this quantity, namely, the force required to extend the spring one unit of length.

TABLE 7.1 Measured lengths of a spring under various loads

Weight (lb) *x*	Measured Length (in.) *y*	Weight (lb) *x*	Measured Length (in.) *y*
0.0	5.06	2.0	5.40
0.2	5.01	2.2	5.57
0.4	5.12	2.4	5.47
0.6	5.13	2.6	5.53
0.8	5.14	2.8	5.61
1.0	5.16	3.0	5.59
1.2	5.25	3.2	5.61
1.4	5.19	3.4	5.75
1.6	5.24	3.6	5.68
1.8	5.46	3.8	5.80

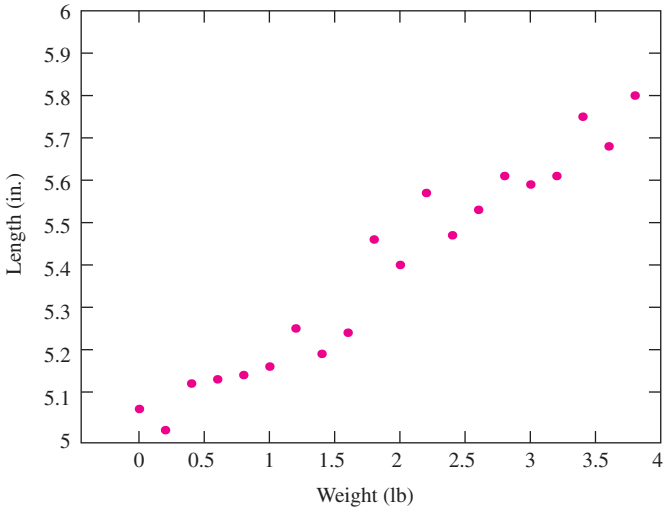

FIGURE 7.9 Plot of measured lengths of a spring versus load.

quantities would be easy to determine. Because of measurement error, β_0 and β_1 cannot be determined exactly, but they can be estimated by calculating the least-squares line.

Figure 7.10 presents the scatterplot of y versus x with the least-squares line superimposed. We write the equation of the line as

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x \tag{7.11}$$

The quantities $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are called the **least-squares coef cients** . The coefficient $\widehat{\beta}_1$, the slope of the least-squares line, is an estimate of the true spring constant β_1, and the coefficient $\widehat{\beta}_0$, the intercept of the least-squares line, is an estimate of the true unloaded length β_0.

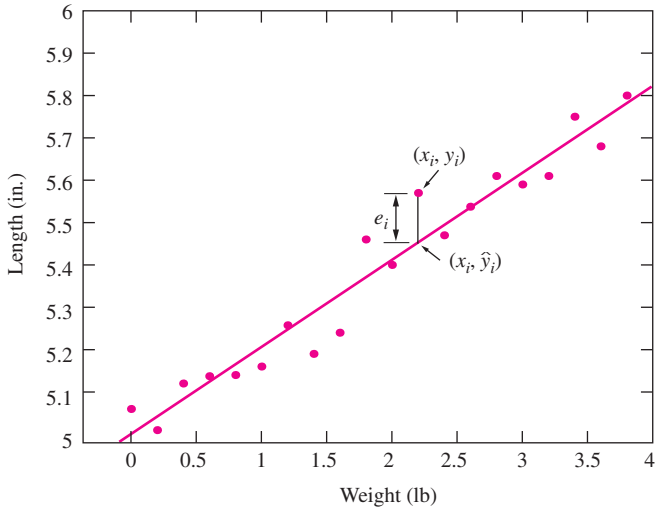

FIGURE 7.10 Plot of measured lengths of a spring versus load. The least-squares line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ is superimposed. The vertical distance from a data point (x_i, y_i) to the point $(x_i, \hat{y}_i)$ on the line is the ith residual e_i. The least-squares line is the line that minimizes the sum of the squared residuals.

The least-squares line is the line that fits the data "best." We now define what we mean by "best." For each data point (x_i, y_i), the vertical distance to the point $(x_i, \hat{y}_i)$ on the least-squares line is $e_i = y_i - \hat{y}_i$ (see Figure 7.10). The quantity $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is called the **tted value**, and the quantity e_i is called the **residual** associated with the point (x_i, y_i). The residual e_i is the difference between the value y_i observed in the data and the fitted value $\hat{y}_i$ predicted by the least-squares line. This is the vertical distance from the point to the line. Points above the least-squares line have positive residuals, and points below the least-squares line have negative residuals. The closer the residuals are to 0, the closer the fitted values are to the observations and the better the line fits the data. We define the least-squares line to be the line for which the sum of the squared residuals $\sum_{i=1}^{n} e_i^2$ is minimized. In this sense, the least-squares line fits the data better than any other line.

In the Hooke's law example, there is only one independent variable (weight), since it is reasonable to assume that the only variable affecting the length of the spring is the amount of weight hung from it. In other cases, we may need to use several independent variables. For example, to predict the yield of a certain crop, we might need to know the amount of fertilizer used, the amount of water applied, and various measurements of chemical properties of the soil. Linear models like Hooke's law, with only one independent variable, are known as **simple linear regression** models. Linear models with more than one independent variable are called **multiple regression** models. This chapter covers simple linear regression. Multiple regression is covered in Chapter 8.

Computing the Equation of the Least-Squares Line

To compute the equation of the least-squares line, we must determine the values for the slope $\widehat{\beta}_1$ and the intercept $\widehat{\beta}_0$ that minimize the sum of the squared residuals $\sum_{i=1}^{n} e_i^2$. To do this, we first express e_i in terms of $\widehat{\beta}_0$ and $\widehat{\beta}_1$:

$$e_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \tag{7.12}$$

Therefore $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the quantities that minimize the sum

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \tag{7.13}$$

These quantities are

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{7.14}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \tag{7.15}$$

Derivations of these results are provided at the end of this section.

Computing Formulas

The quantities $\sum_{i=1}^{n} (x_i - \overline{x})^2$ and $\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$ need to be computed in order to determine the equation of the least-squares line, and as we will soon see, the quantity $\sum_{i=1}^{n} (y_i - \overline{y})^2$ needs to be computed in order to determine how well the line fits the data. When computing these quantities by hand, there are alternate formulas that are often easier to use. They are given in the following box.

Computing Formulas

The expressions on the right are equivalent to those on the left, and are often easier to compute:

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \tag{7.16}$$

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 \tag{7.17}$$

$$\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y} \tag{7.18}$$

Example 7.6

Using the Hooke's law data in Table 7.1, compute the least-squares estimates of the spring constant and the unloaded length of the spring. Write the equation of the least-squares line.

Solution

The estimate of the spring constant is $\widehat{\beta}_1$, and the estimate of the unloaded length is $\widehat{\beta}_0$. From Table 7.1 we compute:

$$\overline{x} = 1.9000 \qquad \overline{y} = 5.3885$$

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = 26.6000$$

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y} = 5.4430$$

Using Equations (7.14) and (7.15), we compute

$$\widehat{\beta}_1 = \frac{5.4430}{26.6000} = 0.2046$$

$$\widehat{\beta}_0 = 5.3885 - (0.2046)(1.9000) = 4.9997$$

The equation of the least-squares line is $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$. Substituting the computed values for $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we obtain

$$y = 4.9997 + 0.2046x$$

Using the equation of the least-squares line, we can compute the fitted values $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ and the residuals $e_i = y_i - \widehat{y}_i$ for each point (x_i, y_i) in the Hooke's law data set. The results are presented in Table 7.2 (page 528). The point whose residual is shown in Figure 7.10 is the one where $x = 2.2$.

In the Hooke's law example, the quantity $\beta_0 + \beta_1 x$ represents the true length of the spring under a load x. Since $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are estimates of the true values β_0 and β_1, the quantity $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ is an estimate of $\beta_0 + \beta_1 x$. Examples 7.7 and 7.8 illustrate this.

Example 7.7

Using the Hooke's law data, estimate the length of the spring under a load of 1.3 lb.

Solution

In Example 7.6, the equation of the least-squares line was computed to be $y = 4.9997 + 0.2046x$. Using the value $x = 1.3$, we estimate the length of the spring under a load of 1.3 lb to be

$$\widehat{y} = 4.9997 + (0.2046)(1.3) = 5.27 \text{ in.}$$

TABLE 7.2 Measured lengths of a spring under various loads, with fitted values and residuals

Weight x	Measured Length y	Fitted Value $\widehat{y}$	Residual e	Weight x	Measured Length y	Fitted Value $\widehat{y}$	Residual e
0.0	5.06	5.00	0.06	2.0	5.40	5.41	−0.01
0.2	5.01	5.04	−0.03	2.2	5.57	5.45	0.12
0.4	5.12	5.08	0.04	2.4	5.47	5.49	−0.02
0.6	5.13	5.12	0.01	2.6	5.53	5.53	−0.00
0.8	5.14	5.16	−0.02	2.8	5.61	5.57	0.04
1.0	5.16	5.20	−0.04	3.0	5.59	5.61	−0.02
1.2	5.25	5.25	0.00	3.2	5.61	5.65	−0.04
1.4	5.19	5.29	−0.10	3.4	5.75	5.70	0.05
1.6	5.24	5.33	−0.09	3.6	5.68	5.74	−0.06
1.8	5.46	5.37	0.09	3.8	5.80	5.78	0.02

Example 7.8

Using the Hooke's law data, estimate the length of the spring under a load of 1.4 lb.

Solution

The estimate is $\widehat{y} = 4.9997 + (0.2046)(1.4) = 5.29$ in.

In Example 7.8, note that the measured length at a load of 1.4 was 5.19 in. (see Table 7.2). But the least-squares estimate of 5.29 in. is based on all the data and is more precise (has smaller uncertainty). We will learn how to compute uncertainties for the estimates $\widehat{y}$ in Section 7.3.

The Estimates Are Not the Same as the True Values

It is important to understand the difference between the least-squares *estimates* $\widehat{\beta}_0$ and $\widehat{\beta}_1$, and the *true values* β_0 and β_1. The true values are constants whose values are unknown. The estimates are quantities that are computed from the data. We may use the estimates as approximations for the true values.

In principle, an experiment such as the Hooke's law experiment could be repeated many times. The true values β_0 and β_1 would remain constant over the replications of the experiment. But each replication would produce different data, and thus different values of the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Therefore $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are *random variables*, since their values vary from experiment to experiment. To make full use of these estimates, we will need to be able to compute their standard deviations. We will discuss this topic in Section 7.3.

The Residuals Are Not the Same as the Errors

A collection of points $(x_1, y_1), \ldots, (x_n, y_n)$ follows a linear model if the x and y coordinates are related through the equation $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. It is important to understand the difference between the residuals e_i and the errors ε_i. Each residual e_i is the difference

$y_i - \widehat{y}_i$ between an observed, or measured, value y_i and the fitted value $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ estimated from the least-squares line. Since the values y_i are known and the values $\widehat{y}_i$ can be computed from the data, the residuals can be computed. In contrast, the errors ε_i are the differences between the y_i and the values $\beta_0 + \beta_1 x_i$. Since the true values β_0 and β_1 are unknown, the errors are unknown as well. Another way to think of the distinction is that the residuals are the vertical distances from the observed values y_i to the least-squares line $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$, and the errors are the distances from the y_i to the true line $y = \beta_0 + \beta_1 x$.

Summary

Given points $(x_1, y_1), \ldots, (x_n, y_n)$:

- The least-squares line is $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$.
- $\widehat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
- $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$
- The quantities $\widehat{\beta}_0$ and $\widehat{\beta}_1$ can be thought of as estimates of a true slope β_1 and a true intercept β_0.
- For any x, $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ is an estimate of the quantity $\beta_0 + \beta_1 x$.

Don't Extrapolate Outside the Range of the Data

What if we wanted to estimate the length of the spring under a load of 100 lb? The least-squares estimate is $4.9997 + (0.2046)(100) = 25.46$ in. Should we believe this? No. None of the weights in the data set were this large. It is likely that the spring would be stretched out of shape, so Hooke's law would not hold. For many variables, linear relationships hold within a certain range, but not outside it. If we extrapolate a least-squares line outside the range of the data, therefore, there is no guarantee that it will properly describe the relationship. If we want to know how the spring will respond to a load of 100 lb, we must include weights of 100 lb or more in the data set.

Summary

Do not extrapolate a fitted line (such as the least-squares line) outside the range of the data. The linear relationship may not hold there.

Don't Use the Least-Squares Line When the Data Aren't Linear

In Section 7.1, we learned that the correlation coefficient should be used only when the relationship between x and y is linear. The same holds true for the least-squares line. When the scatterplot follows a curved pattern, it does not make sense to summarize it with a straight line. To illustrate this, Figure 7.11 (page 530) presents a plot of the