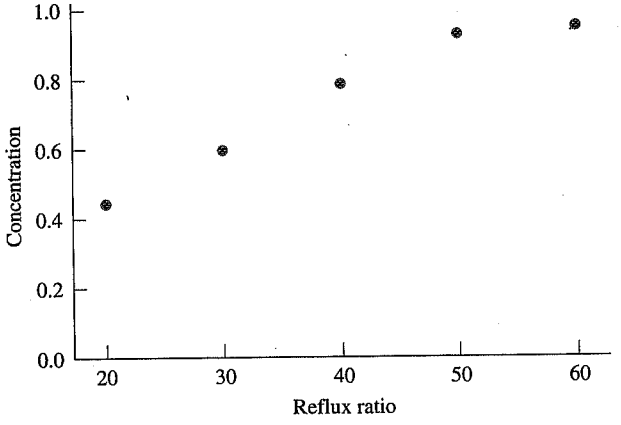
TABLE 6.3
The Reflux Ratio Data

Reflux Ratio	Concentration of Ethanol
20	0.446
30	0.601
40	0.786
50	0.928
60	0.950

FIGURE 6.1
The Scatter Plot for the Reflux Ratio Data



Least Squares Estimation

The Simple Linear Regression Model

Often we can model the response as a straight line in the regressor. Let y_i be the ith response, and let x_i be the ith value for the regressor. The *simple linear regression model* is a linear relationship of the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where β_0 is the y-intercept, β_1 is the slope, and ϵ_i is a random error. Usually we assume that the random errors are independent with mean 0 and variance σ^2. Then the expected value of the response for any value of the regressor is

$$E[y] = \beta_0 + \beta_1 x_i.$$

This equation provides a basis for predicting the response given a specific value for the regressor.

The slope, β_1, represents the expected change in the response given a one-unit change in the regressor and determines the nature of the relationship between the

response and the regressor. If $\beta_1 = 0$, then the response really does not depend on the regressor at all. The expected value of the response does not change as we change the values of the regressor, and we say that the response and the regressor are *uncorrelated.* If $\beta_1 < 0$, then the values of the response grow smaller as we increase the values of the regressor, and we say that the response and the regressor are *negatively correlated.* Conversely, if $\beta_1 > 0$, then the values of the response grow larger as we increase the values of the regressor, and we say that the response and the regressor are *positively correlated.* We should not confuse this sense of correlation with *causality,* which is necessary correlation and is an extremely strong claim about the nature of the relationship between the response and the regressor. Statistics can never establish the necessity of the relationship. Our models show only that the regressor and the response are related, not that one necessarily causes the other.

Our prediction equation depends on the y-intercept, β_0, and the slope, β_1, which in turn are parameters of the population and typically unknown. Traditionally, we estimate these parameters by the *method of least squares,* which minimizes in some meaningful way the errors in predicting the observed data.

A Measure of Overall Fit

The least squares estimates of β_0 and β_1 specifically minimize the *sum of the squares of the residuals.* Let $\hat{y}_i$ be the predicted value of the ith response. If b_0 is our estimate of the y-intercept, β_0, and if b_1 is our estimate of the slope, β_1, then the prediction equation is

$$\hat{y}_i = b_0 + b_1 x_i.$$

An appropriate measure of the quality of the fit of our model looks at the *residuals,* or the differences between the observed values for the response and the predicted values. We may view the residual as an estimate of the error for our prediction of the actual value. Let e_i be the ith residual defined by

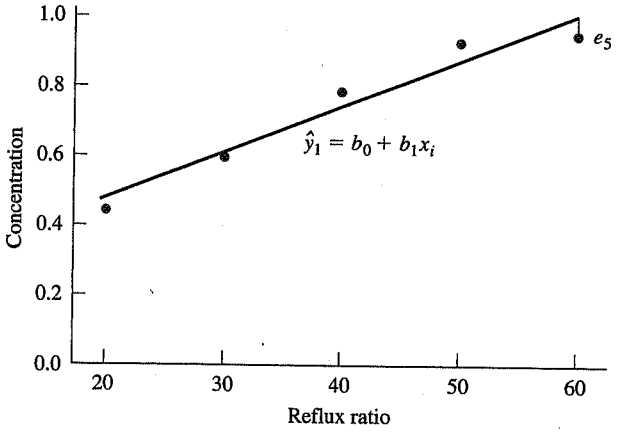
$$e_i = y_i - \hat{y}_i.$$

Figure 6.2 illustrates the residual for the fifth pair of the reflux ratio data. The straight line represents the predicted values. The reflux ratio for the fifth pair is 60. The vertical distance from the straight line to the observed data value is the residual. In this case, since the actual concentration (y_5) is less than the value predicted by the straight line, the residual is negative. Values of e_i near zero, such as the one for a reflux ratio of 30, indicate a good fit.

In the spirit of the sample variance, an appropriate measure of the quality of the fit is the sum of squares for the residuals, SS_{res}, defined by

$$\begin{aligned} SS_{res} &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2. \end{aligned}$$

FIGURE 6.2
Illustration of a Residual



Later we shall see that SS_{res} provides a basis for estimating σ^2, which is the variance of the random errors.

Derivation of the Estimates

We call b_0 and b_1 the *least squares estimators* of β_0 and β_1 if they minimize SS_{res}. By minimizing SS_{res}, b_0 and b_1, in some sense, provide the best straight line equation to fit the data. From calculus, b_0 and b_1 minimize SS_{res} if they satisfy these equations:

$$\frac{\partial}{\partial b_0} \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 = 0$$

$$\frac{\partial}{\partial b_1} \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 = 0.$$

These derivatives result in the following "normal" equations:

$$nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

Solving these two equations simultaneously, we get

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}},$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2.$$

Statistical software packages, spreadsheet programs, and many calculators find these estimates directly. Only rarely do engineers find these estimates by hand. Nonetheless, we can gain some insights into the nature of these estimates from these formulas.

EXAMPLE 6.4 Distillation Column Data—Estimating the Model

For the data in Table 6.4,

$$\sum_{i=1}^{n} x_i = 200$$

$$\sum_{i=1}^{n} x_i^2 = 9000$$

$$\sum_{i=1}^{n} y_i = 3.711$$

$$\sum_{i=1}^{n} x_i y_i = 161.79.$$

TABLE 6.4

Reflux Ratio x_i	Concentration y_i	$x_i y_i$
20	0.446	8.92
30	0.601	18.03
40	0.786	31.44
50	0.928	46.40
60	0.950	57.00