Residuals and the Least-Squares Property

We stated that the regression equation represents the straight line that "best" fits the data. The criterion to determine the line that is better than all others is based on the vertical distances between the original data points and the regression line. Such distances are called *residuals*.

DEFINITION For a pair of sample x and y values, the **residual** is the difference between the *observed* sample value of y and the y value that is *predicted* by using the regression equation. That is,

$$\text{residual} = \text{observed } y - \text{predicted } y = y - \hat{y}$$

x	1	2	4	5
y	4	24	8	32

The paired data are plotted as blue points in Figure 10-6.

So far, this definition hasn't yet won any prizes for simplicity, but you can easily understand residuals by referring to Figure 10-6, which corresponds to the paired sample data shown in the margin. In Figure 10-6, the residuals are represented by the dashed lines.

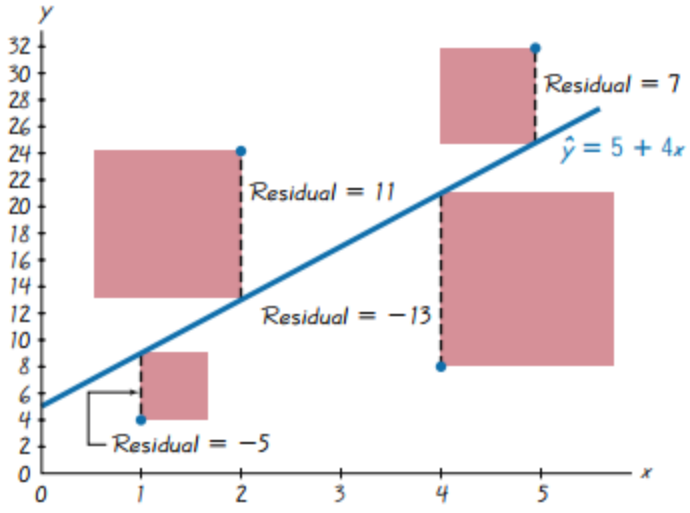

Figure 10-6 Residuals and Squares of Residuals

Consider the sample point with coordinates of (5, 32). If we substitute $x = 5$ into the regression equation $\hat{y} = 5 + 4x$, we get a predicted value of $\hat{y} = 25$. But the actual observed sample value is $y = 32$. The difference $y - \hat{y} = 32 - 25 = 7$ is a residual.

The regression equation represents the line that "best" fits the points according to the following least-squares property.

DEFINITION A straight line satisfies the **least-squares property** if the sum of the squares of the residuals is the smallest sum possible.

From Figure 10-6, we see that the residuals are −5, 11, −13, and 7, so the sum of their squares is

$$(-5)^2 + 11^2 + (-13)^2 + 7^2 = 364$$

We can visualize the least-squares property by referring to Figure 10-6, where the squares of the residuals are represented by the red-square areas. The sum of the red-square areas is 364, which is the smallest sum possible. Use any other straight line, and the red squares will combine to produce an area larger than the combined red area of 364.