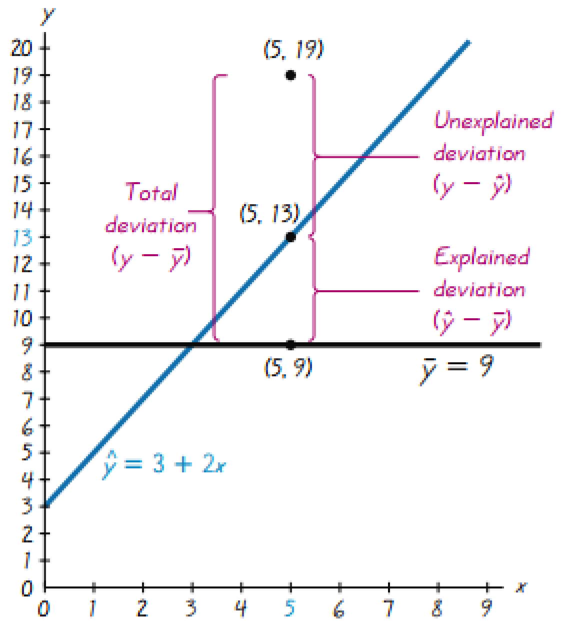
Figure 10-7
Unexplained, Explained, and Total Deviation



Explained and Unexplained Variation

Assume that we have a sample of paired data having the following properties shown in Figure 10-7:

- There is sufficient evidence to support the claim of a linear correlation between x and y.
- The equation of the regression line is $\hat{y} = 3 + 2x$.
- The mean of the y values is given by $\bar{y} = 9$.
- One of the pairs of sample data is $x = 5$ and $y = 19$.
- The point (5, 13) is one of the points on the regression line, because substituting $x = 5$ into the regression equation of $\hat{y} = 3 + 2x$ yields $\hat{y} = 13$.

Figure 10-7 shows that the point (5, 13) lies on the regression line, but the point (5, 19) from the original data set does not lie on the regression line. If we completely ignore correlation and regression concepts and want to predict a value of y given a value of x and a collection of paired (x, y) data, our best guess would be the mean $\bar{y}$. But in this case there is a linear correlation between x and y, so a better way to predict the value of y when $x = 5$ is to substitute $x = 5$ into the regression equation to get $\hat{y} = 13$. We can explain the discrepancy between $\bar{y} = 9$ and $\hat{y} = 13$ by noting that there is a linear relationship best described by the regression line. Consequently, when $x = 5$, the predicted value of y is 13, not the mean value of 9. For $x = 5$, the predicted value of y is 13, but the observed sample value of y is actually 19. The discrepancy between $\hat{y} = 13$ and $y = 19$ cannot be explained by the regression line, and it is called a *residual* or *unexplained deviation*, which can be expressed in the general format of $y - \hat{y}$.

As in Section 3-3 where we defined the standard deviation, we again consider a *deviation* to be a difference between a value and the mean. (In this case, the mean is $\bar{y} = 9$.) Examine Figure 10-7 carefully and note these specific deviations from $\bar{y} = 9$:

Total deviation (from $\bar{y} = 9$) of the point (5, 19) $= y - \bar{y} = 19 - 9 = 10$

Explained deviation (from $\bar{y} = 9$) of the point (5, 19) $= \hat{y} - \bar{y} = 13 - 9 = 4$

Unexplained deviation (from $\bar{y} = 9$) of the point (5, 19) $= y - \hat{y} = 19 - 13 = 6$

These deviations from the mean are generalized and formally defined as follows.