DEFINITIONS

Assume that we have a collection of paired data containing the sample point (x, y), that $\hat{y}$ is the predicted value of y (obtained by using the regression equation), and that the mean of the sample y values is $\bar{y}$.

The **total deviation** of (x, y) is the vertical distance $y - \bar{y}$, which is the distance between the point (x, y) and the horizontal line passing through the sample mean $\bar{y}$.

The **explained deviation** is the vertical distance $\hat{y} - \bar{y}$, which is the distance between the predicted y value and the horizontal line passing through the sample mean $\bar{y}$.

The **unexplained deviation** is the vertical distance $y - \hat{y}$, which is the vertical distance between the point (x, y) and the regression line. (The distance $y - \hat{y}$ is also called a *residual*, as defined in Section 10-3.)

In Figure 10-7 we can see the following relationship for an individual point (x, y):

$$\begin{array}{ccccc} (\text{total deviation}) & = & (\text{explained deviation}) & + & (\text{unexplained deviation}) \\ (y - \bar{y}) & = & (\hat{y} - \bar{y}) & + & (y - \hat{y}) \end{array}$$

The expression above involves deviations away from the mean, and it applies to any one particular point (x, y). If we sum the squares of deviations using all points (x, y), we get amounts of *variation*. The same relationship applies to the sums of squares shown in Formula 10-7, even though the expression above is not algebraically equivalent to Formula 10-7. In Formula 10-7, the **total variation** is the sum of the squares of the total deviation values, the **explained variation** is the sum of the squares of the explained deviation values, and the **unexplained variation** is the sum of the squares of the unexplained deviation values.

Formula 10-7

$$\begin{array}{cccccc} & (\text{total variation}) & = & (\text{explained variation}) & + & (\text{unexplained variation}) \\ \text{or} & \Sigma(y - \bar{y})^2 & = & \Sigma(\hat{y} - \bar{y})^2 & + & \Sigma(y - \hat{y})^2 \end{array}$$

Coefficient of Determination

In Section 10-2 we saw that the linear correlation coefficient r can be used to find the proportion of the total variation in y that can be explained by the linear correlation. This statement was made in Section 10-2:

> **The value of r^2 is the proportion of the variation in y that is explained by the linear relationship between x and y.**

This statement about the explained variation is formalized with the following definition.

DEFINITION The **coefficient of determination** is the proportion of the variation in y that is explained by the regression line. It is computed as

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

We can compute r^2 by using the definition just given with Formula 10-7, or we can simply square the linear correlation coefficient r. Go with squaring r.