## Predicting Condo Prices

A massive study involved 99,491 sales of condominiums and cooperatives in Manhattan. The study used 41 different variables used to predict the value of the condo or co-op. The variables include condition of the unit, the neighborhood, age, size, and whether there are doormen. Some conclusions: With all factors equal, a condo is worth 15.5% more than a co-op; a fireplace increases the value of a condo 9.69% and it increases the value of a co-op 11.36%; an additional bedroom in a condo increases the value by 7.11% and it increases the value in a co-op by 18.53%. This use of statistical methods allows buyers and sellers to estimate value with much greater accuracy. Methods of multiple regression are used when there is more than one predictor variable, as in this study. (Based on data from "So How Much Is That Worth," by Dennis Hevesi, *New York Times*.)

If a multiple regression equation fits the sample data well, it can be used for predictions. For example, if we determine that the multiple regression equation in Example 1 is suitable for predictions, we can use the heights of a mother and father to predict the height of a daughter. But how do we determine whether the multiple regression equation fits the sample data well? Two very helpful tools are the values of adjusted $R^2$ and the $P$-value.

## $R^2$ and Adjusted $R^2$

$R^2$ denotes the **multiple coefficient of determination,** which is a measure of how well the multiple regression equation fits the sample data. A perfect fit would result in $R^2 = 1$, and a very good fit results in a value near 1. A very poor fit results in a value of $R^2$ close to 0. The value of $R^2 = 67.5$ in the Minitab display for Example 1 indicates that 67.5% of the variation in heights of daughters can be explained by the heights of the mothers and fathers. However, the multiple coefficient of determination $R^2$ has a serious flaw: As more variables are included, $R^2$ increases. ($R^2$ could remain the same, but it usually increases.) The largest $R^2$ is obtained by simply including *all* of the available variables, but the best multiple regression equation does not necessarily use all of the available variables. Because of that flaw, it is better to use the *adjusted coefficient of determination,* which is $R^2$ adjusted for the number of variables and the sample size.

**DEFINITION** The **adjusted coefficient of determination** is the multiple coefficient of determination $R^2$ modified to account for the number of variables and the sample size. It is calculated by using Formula 10-8.

**Formula 10-8**

$$\text{adjusted } R^2 = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

where

$n$ = sample size

$k$ = number of predictor ($x$) variables

The preceding Minitab display for the data shows the adjusted coefficient of determination as R-Sq(adj) = 63.7%. If we use Formula 10-8 with the $R^2$ value of 0.675, $n = 20$ and $k = 2$, we find that the adjusted $R^2$ value is 0.637, confirming Minitab's displayed value of 63.7%. When we compare this multiple regression equation to others, it is better to use the adjusted $R^2$ of 63.7% (or 0.637).

## $P$-Value

The $P$-value is a measure of the overall significance of the multiple regression equation. The displayed Minitab $P$-value of 0.000 (rounded to three decimal places) is small, indicating that the multiple regression equation has good overall significance and is usable for predictions. That is, it makes sense to predict heights of daughters based on heights of mothers and fathers. Like the adjusted $R^2$, this $P$-value is a good measure of how well the equation fits the sample data. The value of 0.000 results from a test of the null hypothesis that $\beta_1 = \beta_2 = 0$. Rejection of $\beta_1 = \beta_2 = 0$ implies that at least one of $\beta_1$ and $\beta_2$ is not 0, indicating that this regression equation