is effective in predicting heights of daughters. A complete analysis of the Minitab results might include other important elements, such as the significance of the individual coefficients, but we will keeps things simple (!) by limiting our discussion to the three key components—multiple regression equation, adjusted R^2, and P-value.

Finding the Best Multiple Regression Equation

When trying to find the best multiple regression equation, we should not necessarily include all of the available predictor variables. Finding the best multiple regression equation requires abundant use of judgment and common sense, and there is no exact and automatic procedure that can be used to find the best multiple regression equation. *Determination of the best multiple regression equation is often quite difficult and is beyond the scope of this section*, but the following guidelines are helpful.

Guidelines for Finding the Best Multiple Regression Equation

1. *Use common sense and practical considerations to include or exclude variables.* For example, when trying to find a good multiple regression equation for predicting the height of a daughter, we should exclude the height of the physician who delivered the daughter, because that height is obviously irrelevant.
2. *Consider the P-value.* Select an equation having overall significance, as determined by the P-value found in the computer display.
3. *Consider equations with high values of adjusted R^2, and try to include only a few variables.* Instead of including almost every available variable, try to include relatively few predictor (x) variables. Use these guidelines:
   - Select an equation having a value of adjusted R^2 with this property: If an additional predictor variable is included, the value of adjusted R^2 does not increase very much.
   - For a particular number of predictor (x) variables, select the equation with the largest value of adjusted R^2.
   - In excluding predictor (x) variables that don't have much of an effect on the response (y) variable, it might be helpful to find the linear correlation coefficient r for each pair of variables being considered. If two predictor values have a very high linear correlation coefficient, there is no need to include them both, and we should exclude the variable with the lower value of r.

The following example illustrates that common sense and *critical thinking* are essential tools for effective use of methods of statistics.

Example 2 Predicting Height from Footprint Evidence

Data Set 2 in Appendix B includes the age, foot length, shoe print length, shoe size, and height for each of 40 different subjects. Using those sample data, find the regression equation that is best for predicting height. Is the best regression equation a *good* equation for predicting height?

Solution

Using the response variable of height and possible predictor variables of age, foot length, shoe print length, and shoe size, there are 15 different possible combinations of predictor variables. Table 10-5 includes key results from five of those combinations. Blind and thoughtless application of regression methods would suggest that the best regression equation uses all four of the predictor variables, because that combination yields the highest adjusted R^2 value of 0.7585. However, given

Clinical Trial Cut Short

What do you do when you're testing a new treatment and, before your study ends, you find that it is clearly effective? You should cut the study short and inform all participants of the treatment's effectiveness. This happened when hydroxy-urea was tested as a treatment for sickle cell anemia. The study was scheduled to last about 40 months, but the effectiveness of the treatment became obvious and the study was stopped after 36 months. (See "Trial Halted as Sickle Cell Treatment Proves Itself," by Charles Marwick, *Journal of the American Medical Association*, Vol. 273, No. 8.)