DEFINITION A **multiple regression equation** expresses a linear relationship between a response variable y and two or more predictor variables ($x_1, x_2, \ldots, x_k$). The general form of a multiple regression equation obtained from sample data is

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

The accompanying box includes the key elements of this section. For notation, the coefficients $b_0, b_1, b_2, \ldots, b_k$ are sample *statistics* used to estimate the corresponding population parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$. Also, note that the multiple regression equation is a natural extension of the format $\hat{y} = b_0 + b_1x_1$ used in Section 10-3 for regression equations with a single independent variable x_1. In Section 10-3, it would have been reasonable to question why we didn't use the more common format of $y = mx + b$, and we can now see that using $\hat{y} = b_0 + b_1x_1$ allows us easily to extend that format to include additional predictor variables, as shown below.

Finding a Multiple Regression Equation

Objective

Use sample data from three or more variables to find a multiple regression equation that is useful for predicting values of the response variable y.

Notation

$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$ (Multiple regression equation found from *sample* data)

$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots + \beta_kx_k$ (Multiple regression equation for the *population* of data)

$\hat{y}$ = predicted value of y (computed using the multiple regression equation)

k = number of *predictor* variables (also called *independent variables* or x variables)

n = sample size (number of values for any one of the variables)

Requirements

For any specific set of x values, the regression equation is associated with a random error often denoted by ε. We assume that such errors are normally distributed with a mean of 0 and a standard deviation of σ and that the random errors are independent. Because these assumptions are difficult to check, we assume throughout this section that the necessary requirements are satisfied.

Procedure for Finding a Multiple Regression Equation

Manual calculations are not practical, so computer software such as STATDISK, Minitab, Excel, StatCrunch, or a TI-83/84 Plus calculator must be used. (See the "Using Technology" instructions at the end of this section.)

Because we use technology to find multiple regression equations, we ignore the actual calculations and focus instead on *interpreting* computer displays.

In 1886, Francis Galton was among the first to study genetics using the methods of regression we are now considering. He wrote the article "Regression Towards Mediocrity in Hereditary Stature" claiming that heights of offspring regress or revert back toward a mean. Example 1 involves that same historical context with more recent data.