Multiple regression provides a method of predicting a response variable $y$ from two or more explanatory $x$ variables.

Multiple regression is potentially much more useful than simple regression, because there are often several important explanatory factors, instead of just one.

The purposes of multiple regression are the same as in simple regression:
(1) Describing the relationship.

- Salary \((y)\) of company employees may depend on several factors, such as years of experience \((x_1)\), years of education \((x_2)\), and gender \((x_3\), represented as 0 or 1 to distinguish male and female). 

Describing and understanding how these \(x\) factors influence \(y\) would provide important evidence in a gender discrimination lawsuit.

The regression coefficient for gender would give an estimate of how large the salary gap is for men and women *after adjustment* for age and experience.
(2) Predicting a new observation.

- Suppose we study the finishing time \( (y) \) for horses racing at Belmont Park. If we know how \( y \) depends on the length of the race \( (x_1) \), the post position \( (x_2) \), the horse’s finishing time in his previous race \( (x_3) \), the length of his previous race \( (x_4) \), his total life time winnings \( (x_5) \) and other factors, we could try to forecast the winning times of all horses in tomorrow’s race, and thereby forecast the winner.

(3) Adjusting and controlling a process.

- If we know how ice cream sales \( (y) \) depend on fat content \( (x_1) \), price \( (x_2) \), sugar content \( (x_3) \), and other factors, we can reformulate our product and set its price to (hopefully) increase profits.
The Model

Suppose we have \( p \) explanatory variables \( x_1, \ldots, x_p \) which we feel may be useful in predicting \( y \). Suppose that for each observation \( y_i \) of the response variable, we have the corresponding observations \( x_{i1}, \ldots, x_{ip} \) on the explanatory variables.

The multiple linear regression model is

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + u_i
\]

\((i=1, \ldots, n)\), where \( \beta_0, \ldots, \beta_p \) are unknown (but nonrandom) parameters and \( u_i \) are iid random errors, independent of all explanatory variables, with \( E[u_i] = 0, \ Var[u_i] = \sigma_u^2 \), a constant, which does not change with the explanatory variables.
The error term $u_i$ is included to take account of the randomness in our response $y_i$.

The **true regression function**, or **response surface**,

$$E(y \mid x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

accounts for the part of $y$ which depends systematically on the explanatory variables. Since there are several explanatory variables, the response surface is not a line, but a flat surface called a hyperplane.

So the observed response $y$ is the *sum* of a systematic part (the response surface) and a random part (the error). We seek to estimate the response surface based on our data. This is usually done by
least squares, i.e., so as to make the residual sum of squares as small as possible. Given actual data, the least squares estimates can be obtained by using Minitab, or some other software. We will give formulas later.

**Interpreting The Coefficients**

Suppose we want to predict the selling price ($y$; Thousands of Dollars) of a single family home based on the house size ($x_1$; Hundreds of Square Feet), the age ($x_2$; Years), and the lot size ($x_3$; Thousands of Square Feet).

The estimated regression coefficients are

\[
b_0 = -16.06, \ b_1 = 4.146, \ b_2 = -0.236, \ b_3 = 4.831.
\]
The fitted response surface is therefore

$$\hat{y} = -16.06 + 4.146 x_1 - .236 x_2 + 4.831 x_3 .$$

**Interpretation of the Regression Coefficients:** The estimated regression coefficient for a given $x$-variable estimates the effect of that $x$ variable on $E[y]$ *after an adjustment has been made for the other $x$-variables*. In other words, the coefficient $b_j$ (for the $j$’th $x$-variable, $x_j$) estimates how much larger you expect $y$ to be if $x_j$ is increased by one unit and *all other $x$-variables are held fixed*. Note that the meaning and interpretation of a given regression coefficient depends on what other variables are included in the model.
Eg: In the above example, it is not wise to try to interpret the term $b_0$ since we have no data for $x_1$, $x_2$, $x_3$ near zero. Nevertheless, it is almost always a good idea to include a constant term ($\beta_0$) in the model, as this will usually produce better predictions.

Since we obtained $b_2 = -0.236$, we can say that for a given house size and a given lot size, we estimate that each year of age reduces the mean selling price by $236$. If we run a simple regression of $y$ using the single explanatory variable $x_2$, we find that

$$\hat{y} = 112.47 - 2.7268x_2.$$
Thus, *without* controlling for house size and lot size, we estimate that each additional year of age reduces the mean selling price by an average of $2,727. This is *very different* from the $236 figure obtained above, which adjusts for house size and lot size.

As we see, the interpretation of a regression coefficient, as well as its numerical value, depends on what other variables are used in the model. This is one reason why we need to think carefully about which variables to use.
Using The Multiple Regression Equation

We can estimate the response surface $E(y \mid x)$ and predict a future value using $\hat{y}$. We can also obtain confidence intervals for $E(y \mid x)$ and prediction intervals for a future $y$ from the Minitab output. For example, if a house has a ground area of 2000 feet ($x_1=20$), the house is 10 years old ($x_2=10$) and the lot size is 10,000 square feet ($x_3=10$), then $\hat{y}=112.81$, so the predicted selling price is $112,810. The 95 percent confidence interval for the mean selling price is ($104,940 , 120,690$) and the 95 percent prediction interval for the price of the house is ($95,710 , 129,910$).
Matrix Formulation of the Model

It is very helpful to re-write the multiple linear regression model in matrix form. Define the column vectors \( y = (y_1, \ldots, y_n)' \), \( u = (u_1, \ldots, u_n)' \), and \( \beta = (\beta_0, \ldots, \beta_p)' \). Load the explanatory variables, one column at a time, into an \( n \times (p+1) \) matrix,

\[
X = \begin{bmatrix}
1 & x_{11} & \cdots & x_{1p} \\
1 & x_{21} & \cdots & x_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & \cdots & x_{np}
\end{bmatrix}
\]

We will refer to the columns of this matrix as \( x_0, \ldots, x_p \). So \( x_0 \) is a column of ones, \( x_1 = (x_{11}, x_{21}, \ldots, x_{n1})' \) is a column vector of the \( n \) observations on the first explanatory variable, etc.
Then the multiple linear regression model can be written as

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n
\end{bmatrix} = 
\begin{bmatrix}
  1 & x_{11} & \cdots & x_{1p} \\
  1 & x_{21} & \cdots & x_{2p} \\
  \vdots & \vdots & \ddots & \vdots \\
  1 & x_{n1} & \cdots & x_{np}
\end{bmatrix} 
\begin{bmatrix}
  \beta_0 \\
  \beta_1 \\
  \vdots \\
  \beta_p
\end{bmatrix} + 
\begin{bmatrix}
  u_1 \\
  u_2 \\
  \vdots \\
  u_n
\end{bmatrix},
\]

or, in matrix notation,

\[
y = X\beta + u.
\]