BONUS
CHAPTER

3

Correlation and Simple Linear Regression

For the last several chapters, we have put inferential statistics to work drawing conclusions about one, two, or more population means and proportions. I know this has been a lot of fun for you, but it's time to move to another type of inferential statistics that is even more exciting. (If you can imagine that!)

This final chapter focuses on describing how two variables relate to one another. Using correlation and simple regression, we will be able to first determine whether a relationship does indeed exist between the variables and second describe the nature of this relationship in mathematical terms. And hopefully we'll have some fun doing it!

In This Chapter

- Determining the correlation between two variables and performing a simple linear regression
- Calculating a confidence interval for a regression line
- Performing a hypothesis test on the coefficient of the regression line
- Using Excel to calculate the correlation coefficient and perform simple linear regression

Correlation Coefficient

Correlation measures both the strength and direction of the relationship between two variables, *x* and *y*. Figure 3.1 illustrates the different types of correlation in a series of scatter plots, which graphs each ordered pair of (*x*,*y*) values. The convention is to place the *x* variable on the horizontal axis and the *y* variable on the vertical axis.

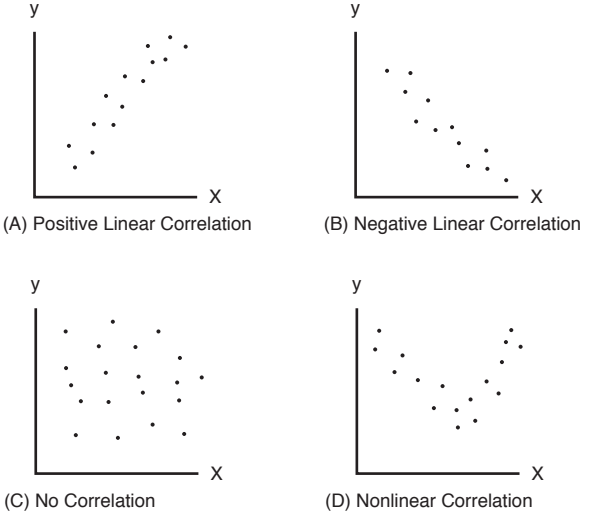

Figure 3.1
Different types of correlation.

Graph A in Figure 3.1 shows an example of positive linear correlation where, as *x* increases, *y* also tends to increase in a linear (straight line) fashion. Graph B shows a negative linear correlation where, as *x* increases, *y* tends to decrease linearly. Graph C indicates no correlation between *x* and *y*. This set of variables appears to have no connection with one another. And finally, Graph D is an example of a nonlinear relationship between variables. As *x* increases, *y* decreases at first and then changes direction and increases.

For the remainder of this chapter, we will focus on linear relationships between the independent and dependent variables. Nonlinear relationships can be very disagreeable and go beyond the scope of this book. Before we start, let's review the independent and dependent variables, which we discussed back in Chapter 2.

Review of Independent and Dependent Variables

Suppose I would like to investigate the relationship between the number of hours that a student studies for a statistics exam and the grade for that exam (uh-oh). The following table shows sample data from six students whom I randomly chose.

Data for Statistics Exam

Hours Studied	Exam Grade
3	86
5	95
4	92
4	83
2	78
3	82

Obviously, we would expect the number of hours studying to affect the grade. The Hours Studied variable is considered the *independent variable* (x) because it explains the observed variation in the Exam Grade, which is considered the *dependent variable* (y). The data from the previous table are considered *ordered pairs* of (x,y) values, such as (3,86) and (5,95).

DEFINITION

The **independent variable** (x) explains the variation in the **dependent variable** (y).

This relationship between the independent and the dependent variables only exists in one direction, as shown here:

Independent variable (x) → Dependent variable (y)

This relationship does not work in reverse. For instance, we would not expect that the Exam Grade variable would explain the variations in the number of hours studied in our previous example.

WRONG NUMBER

Exercise caution when deciding which variable is independent and which is dependent. Examine the relationship from both directions to see which one makes the most sense. The wrong choice will lead to meaningless results.

Other examples of independent and dependent variables are shown in the following table.

Examples of Independent and Dependent Variables

Independent Variable	Dependent Variable
Size of TV	Selling price of TV
Level of advertising	Volume of sales
Size of sports team payroll	Number of games won

Now, let's focus on describing the relationship between the x and y variables using inferential statistics.

Understanding and Calculating the Correlation Coefficient

The sample *correlation coefficient*, r, provides us with both the strength and direction of the relationship between the independent and dependent variables. Values of r range between -1.0 and +1.0. When r is positive, the relationship between x and y is positive (for example, Graph A from Figure 3.1), and when r is negative, the relationship is negative (Graph B). A correlation coefficient close to 0 is evidence that there is no relationship between x and y (Graph C).

DEFINITION

The sample **correlation coefficient**, r, indicates both the strength and direction of the relationship between the independent and dependent variables. Values of r range from –1.0, a strong negative relationship, to +1.0, a strong positive relationship. When $r = 0$, there is no relationship between variables x and y.

The strength of the relationship between *x* and *y* is measured by how close the correlation coefficient is to +1.0 or -1.0 and can be viewed in Figure 3.2.

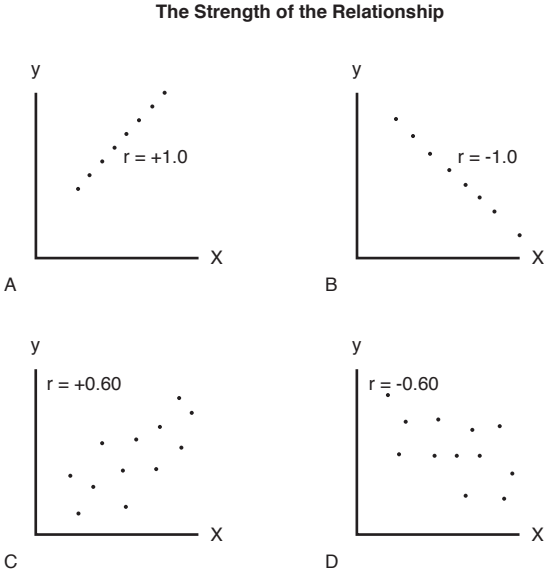

Figure 3.2
The strength of the relationship.

Graph A illustrates a perfect positive correlation between *x* and *y* with *r = +1.0*. Graph B shows a perfect negative correlation between *x* and *y* with *r = -1.0*. Graphs C and D are examples of weaker relationships between the independent and dependent variables.

We can calculate the correlation coefficient using the following equation:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n(\Sigma x^2) - (\Sigma x)^2\right]\left[n(\Sigma y^2) - (\Sigma y)^2\right]}}$$

Wow! I know this looks overwhelming, but before we panic, let's try out our exam grade example on this. The following table will help break down the calculations and make them more manageable.

Hours of Study x	**Exam Grade** y	xy	x^2	y^2
3	86	258	9	7396
5	95	475	25	8464
4	92	368	16	9025
4	83	332	16	6889
2	78	156	4	6084
3	82	246	9	6724
$\sum x = 21$	$\sum y = 516$	$\sum xy = 1{,}835$	$\sum x^2 = 79$	$\sum y^2 = 44{,}582$

Keep these five summation numbers handy as we will use them throughout this chapter. Using these values along with $n = 6$, the sample size, we have:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\left(\Sigma x^2\right) - (\Sigma x)^2\right]\left[n\left(\Sigma y^2\right) - (\Sigma y)^2\right]}}$$

$$r = \frac{6(1{,}835) - (21)(516)}{\sqrt{\left[6(79) - (21)^2\right]\left[6(44{,}582) - (516)^2\right]}} \frac{174}{\sqrt{(33)(1{,}236)}} = 0.862$$

As you can see, we have a fairly strong positive correlation between hours of study and the exam grade. That's good news for us teachers.

What is the benefit of establishing a relationship between two variables such as these? That's an excellent question. When we discover that a relationship does exist, we can predict exam scores based on a particular number of hours of study. Simply put, the stronger the relationship, the more accurate our prediction will be. You will learn how to make such predictions later in this chapter when we discuss simple linear regression.

WRONG NUMBER

Be careful to distinguish between $\sum x^2$ and $\left(\sum x\right)^2$. With $\sum x^2$, we first square each value of x and then add each squared term. With $\left(\sum x\right)^2$, we first add each value of x and then square this result. The answers between the two are very different!

Testing the Significance of the Correlation Coefficient

The correlation coefficient we calculated is based on a sample of data. The population correlation coefficient, denoted by the symbol ρ (a Greek letter pronounced *rho*), measures the correlation between the hours of study and exam grades for all students. Because we only used a sample, not the entire population, we don't know the value of the population correlation coefficient, ρ. We can perform a hypothesis test to determine whether the population correlation coefficient, ρ, is significantly different from 0 based on the value of the calculated sample correlation coefficient, r. We can state the hypotheses as:

$H_0 : \rho \leq 0$

$H_1 : \rho > 0$

This statement tests whether a positive correlation exists between x and y. I could also choose a two-tail test that would investigate whether any correlation exists (either positive or negative) by setting $H_0 : \rho = 0$ and $H_1 : \rho \neq 0$.

The calculated t-test statistic for the correlation coefficient uses the Student's t-distribution as follows:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

where:

r = the sample correlation coefficient

n = the sample size

For the exam grade example, the calculated t-test statistic becomes:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.862\sqrt{\frac{6-2}{1-(0.862)^2}} = 0.862\sqrt{\frac{4}{0.257}} = 3.401$$

The critical t-value is based on $d.f. = n - 2$ if we choose $\alpha = 0.05$ and $t_c = 2.132$ from Table 4 in Appendix B for a one-tail test. Because the calculated t-test statistic $t > t_c$ (the critical value), we reject H_0 and conclude that there is indeed a positive correlation coefficient between hours of study and the exam grade. Once again, statistics has proven that all is right in the world!

Using Excel to Calculate the Correlation Coefficient

After looking at the nasty calculations involved for the correlation coefficient, I'm sure you'll be relieved to know that Excel will do the work for you with the CORREL function that has the following characteristics:

CORREL(array1, array2)

where:

array1 = the range of data for the first variable

array2 = the range of data for the second variable

For instance, Figure 3.3 shows the CORREL function being used to calculate the correlation coefficient for the exam grade example.

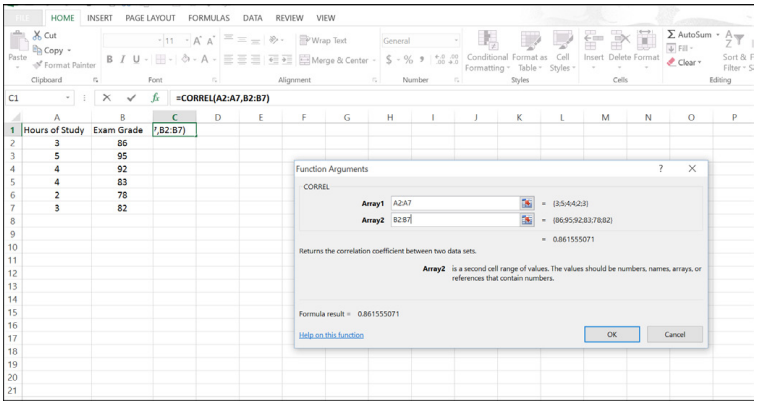

Figure 3.3
CORREL function in Excel with the exam grade example.

Cell C1 contains the Excel formula =CORREL(A2:A7,B2:B7) with the result being 0.862.

Simple Linear Regression

Regression analysis is very useful in any area. It has numerous applications. No matter what your area of study or work is, chances are regression can be very helpful to you. Regression quantifies a relationship between two (or more) variables so we can connect theory to reality. In our previous example, it quantifies the relationship between the hours of study and the exam grade enabling us to predict the average exam grade for a student who studied a specific number of hours.

What Is Simple Linear Regression?

The technique of *simple linear regression* enables us to describe a straight line that best fits the data for our variables x and y. The estimated equation for a straight line, known as a *linear equation,* takes the form:

$$\hat{y}_i = a + b\,x_i$$

where:

$\hat{y}_i$ = the predicted value of y, given a value of x

x_i = the independent variable

a = the y-intercept for the straight line

b = the slope of the straight line

Using data for our variables, x and y, we calculate the values for a and b and place them in the equation instead of a and b.

DEFINITION

The technique of **simple linear regression** enables us to describe a straight line that best fits the data for the x and y variables.

Figure 3.4 illustrates this concept.

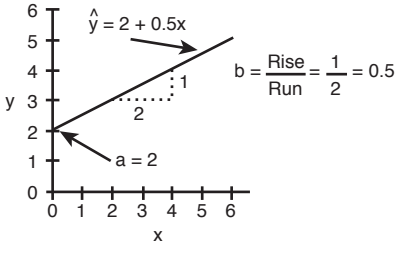

Figure 3.4
Equation for a straight line.

Figure 3.4 shows a line described by the equation $\hat{y}_i = 2 + 0.5\,x_i$. The y-intercept is the point where the line crosses the y-axis, which in this case is $a = 2$. The slope of the line, b, is shown as the ratio of the rise of the line over the run of the line, shown as $b = 0.5$. A positive slope indicates the line is rising from left to right. A negative slope, you guessed it, moves lower from left to right. If $b = 0$, the line is horizontal, which means there is no relationship between the

independent and dependent variables. In other words, a change in the value of x has no effect on the value of y.

Students sometimes struggle with the distinction between $\hat{y}$ and y. Figure 3.5 shows six ordered pairs and a line that appears to fit the data described by the equation $\hat{y}_i = 2 + 0.5\, x_i$.

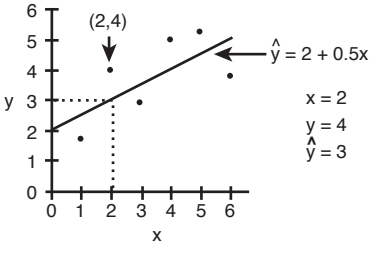

Figure 3.5
The difference between y and $\hat{y}$.

Figure 3.5 shows a data point that corresponds to the ordered pair $x = 2$ and $y = 4$. Notice that the *predicted* value of y according to the line at $x = 2$ is $\hat{y} = 3$. We can verify this using the equation as follows:

$$\hat{y}_i = 2 + 0.5\, x_i = 2 + 0.5(2) = 3$$

The value of y represents an actual data point, while the value of $\hat{y}$ is the predicted value of y using the estimated linear equation, given a value for x.

Our next step is to find the linear equation that best fits the data.

The Ordinary Least Squares Method

The *ordinary least squares method* (*OLS* for short) is a mathematical procedure to identify the linear equation that best fits the data by finding values for a, the y-intercept; and b, the slope. The goal of the ordinary least squares method is to minimize the sum of the squared difference between the values of y and $\hat{y}$. If we define the residuals e_i as $e_i = y_i - \hat{y}_i$, the OLS method will minimize the sum of the squared residuals as follows:

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

Where n is the number of observations.

This concept is illustrated in Figure 3.6.

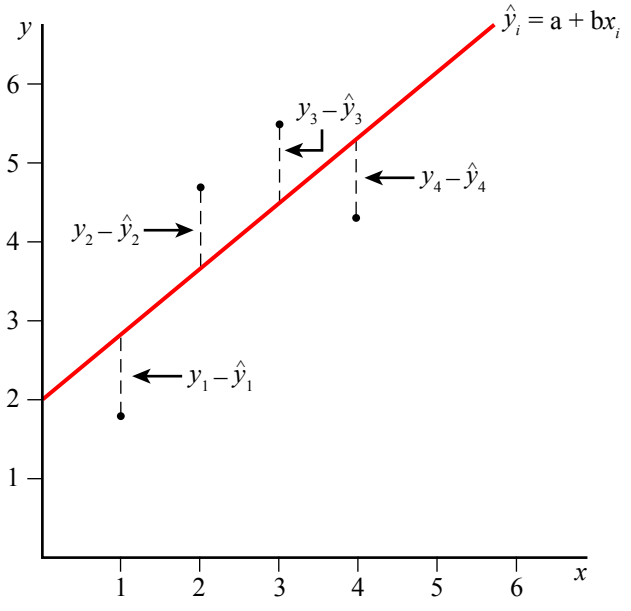

Figure 3.6
Minimizing the residuals.

According to Figure 3.6, the line that best fits the data, the *regression line*, will minimize the total squared residual of each data point. We'll demonstrate how to determine this regression equation using the least squares method through the following example.

DEFINITION

The **ordinary least squares (OLS) method** is a mathematical procedure to identify the linear equation that best fits the data by finding values for *a,* the *y*-intercept; and *b,* the slope. The goal of the least squares method is to minimize the sum of the squared difference between the values of *y* and $\hat{y}$. The **regression line** is the line that best fits the data.

Suppose you and your spouse are planning to buy a house. Your spouse tells you, "Now that you know statistics, why don't you do some analysis on home prices?" You thought about it and decided to run a regression to estimate home prices based on their sizes. So you collect data for the size of the house and its price for 10 houses in the neighborhood you want to live in. You found the following data where size is measured in square feet and price is measure in dollars:

Homes' Sizes and Prices Data

Size	Price	Size	Price
1290	290,000	1200	204,000
1480	280,000	1900	289,900
1660	275,000	1100	198,900
2480	399,900	1600	235,000
2755	515,000	1300	269,000

Because your goal is to determine the price of a house based on its size, the size of the house will be the independent variable and the price of the house will be the dependent variable.

The OLS method finds the linear equation that best fits the data by determining the values for a and b using the following equations:

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$a = \overline{y} - b\overline{x}$$

where:

$\overline{x}$ = the average value of x, the independent variable

$\overline{y}$ = the average value of y, the dependent variable

If you notice the formulas above use the same five summations we used earlier in calculating the correlation coefficient. That's why I told you to keep them handy. The following table, which uses the data for home prices in thousands of dollars to make the calculations easier, summarizes the calculations necessary for these equations.

Calculations for the Slope and Intercept

Size (x)	Price (y)	xy	x^2	y^2
1290	290	374,100	1,664,100	84,100
1480	280	414,400	2,190,400	78,400
1660	275	456,500	2,755,600	75,625
2480	399.9	991,752	6,150,400	159,920
2755	515	1,418,825	7,590,025	265,225

Size (x)	**Price (y)**	xy	x^2	y^2
1200	204	244,800	1,440,000	41,616
1900	289.9	550,810	3,610,000	84,042
1100	198.9	218,790	1,210,000	39,561
1600	235	376,000	2,560,000	55,225
1300	269	349,700	1,690,000	72,361
$\sum x = 16,765$	$\sum y = 2,957$	$\sum xy = 5,395,677$	$\sum x^2 = 30,860,525$	$\sum y^2 = 956,075$

$$\overline{x} = \frac{\Sigma x}{n} = \frac{16,765}{10} = 1,676.5 \qquad \overline{y} = \frac{\Sigma y}{n} = \frac{2,957}{10} = 295.7$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{10(5,395,677) - (16,765)(2,957)}{10(30,860,525) - (16,765)^2} = \frac{4,382,665}{27,540,025} = 0.159138$$

$$a = \overline{y} - b\overline{x} = 295.7 - 0.159138(1,676.5) = 28.905$$

The estimated regression equation for the home prices example would be:

$$\hat{y}_i = 28.905 + 0.15914\, x_i$$

Because the slope of this equation is a positive 0.15914, this means that as the size of the house increases so does the price. In particular, the price of the house will increase by $159 (which is 0.159 x 1000) for every square foot increase in the size of the house.

I can use this estimated equation to predict the price of a house based on its size. For example, if you and your spouse want a 1,200 square feet house, the estimated average price for it will be

$$\hat{y}_i = 28.905 + 0.15914\, x_i = 28.905 + 0.15914\ (1200) = \$219,873$$

So we expect the average price of a 1200 square feet house to be $219,873. If you find a 1,200 square feet house in this neighborhood for $200,000 then it is a good deal!

Measures of Goodness of Fit for the Model

Now, your spouse is questioning your estimate and asking how accurate it is. We have two ways to check how good our estimated regression line is: the coefficient of determination and the standard error of the regression. So let's look at each one of them.

1. The Coefficient of Determination

The *coefficient of determination,* R^2, measures the percentage of the variation in y that is explained by the variation in x. If $R^2 = 1$, all of the variation in y is explained by the variable x. If $R^2 = 0$, none of the variation in y is explained by the variable x.

To understand the coefficient of determination, we need to know a few more concepts: the total variation, the explained variation, and the unexplained variation. The total variation is the difference between the actual value of y and the average value of y as in this equation:

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

This total variation can be divided (or as statisticians call it "partitioned") into two parts: the part that is explained by our regression line and the part that is unexplained by our regression line. The explained variation is the difference between $\hat{y}$ and $\overline{y}$ as in the following equation:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

The unexplained variation is the difference between y and $\hat{y}$ and is calculated as:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

BOB'S BASICS

The explained variation **SSR** is also known as the regression sum of squares, the unexplained variation **SSE** is the error sum of squares, and the total variation **SST** is the total sum of squares.

The explained variation is the part of the total variation that is explained by our regression line and the variable(s) included in it, whereas the unexplained variation is the part that is not explained by our regression line and reflects the effect of other variables that are not included in our model. Both the explained and the unexplained variations add up to the total variation. As a result, the larger the explained variation part is compared to the total, the better our estimated regression line is. To put it differently, the smaller the unexplained variation part is compared to the total variation, the better our estimated regression line is.

To clarify this, let's look at any one data point in our example, say (2755, 515). When $x = 2755$, $y_i = 515$, $\overline{y} = 295.7$ (as calculated above) and $\hat{y}_i = 28.905 + 0.15914(2755) = 467.5$. The total variation, $y_i - \overline{y} = 515 - 295.7 = 219.3$ is divided into the explained variation $\hat{y}_i - \overline{y} = 467.5 - 295.7 = 171.8$ and the unexplained variation $y_i - \hat{y}_i = 515 - 467.5 = 47.5$. So out of \$219.3 thousand, our regression line explained \$171.8 thousand and the error part was only \$47.5 thousand. As you can see from these numbers, the larger the explained variation part is, the better our model is. Figure 3.7 presents these variations.

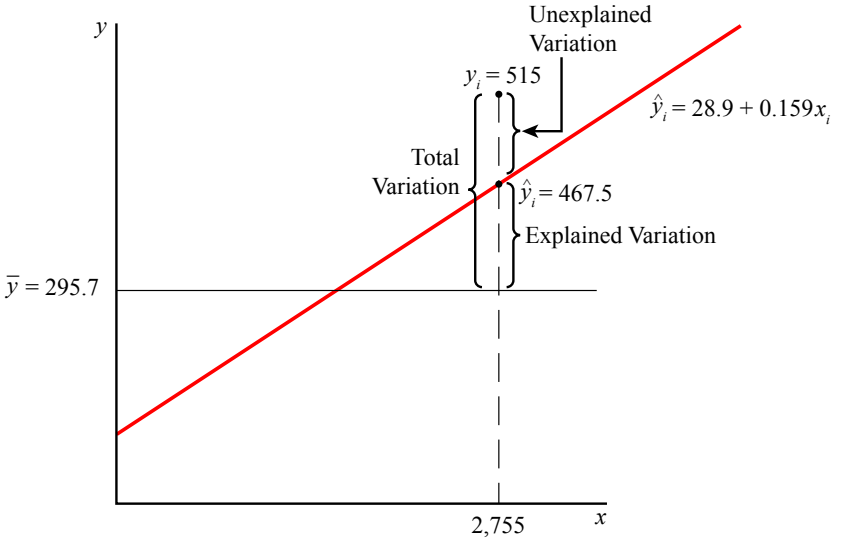

Figure 3.7
Total, explained, and unexplained variations.

The coefficient of determination measures the percentage of the explained variation relative to the total variation and is calculated as:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

To simplify the calculations, we can use the following equations:

$$SST = \sum_{i=1}^{n} y^2 - \frac{\left(\sum_{i=1}^{n} y\right)^2}{n}$$

$$SSE = \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y - b\sum_{i=1}^{n} xy$$

$$SSR = SST - SSE$$

These equations are much easier when they are plugged in with numbers, so let's apply them to our home prices example.

$$SST = 956{,}075 - \frac{(2{,}957)^2}{10} = 81{,}690.10$$

$$SSE = 956{,}075 - 28.905\,(2{,}957) - 0.15914\,(5{,}395{,}677) = 11{,}934.88$$

$$SSR = 81{,}690.10 - 11{,}934.88 = 69{,}755.22$$

The coefficient of determination becomes:

$$R^2 = \frac{SSR}{SST} = \frac{69{,}755.22}{81{,}690.10} = 0.854$$

In other words, in this neighborhood 85.4 percent of the variation in the price of the house is explained by the variation in its size. Now you can see that your estimated regression line is very reliable!

DEFINITION

The **coefficient of determination,** R^2, measures the percentage of the variation in y that is explained by the variation in x.

2. The Standard Error of the Regression

The standard error of the regression measures the amount of dispersion of the observed data around the estimated regression line. If the data points are close to the estimated regression line, then the standard error of the regression is small and vice versa. As with any standard error, the closer the data are to the estimated line, the smaller the standard error of the regression is, and the better our estimated regression line is and vice versa. Panel A of Figure 3.8 shows an estimated line with a smaller standard error of the regression. As you can see, the data points are close to the estimated regression line, whereas in Panel B the data points are further away from the estimated regression line. The *standard error of the regression*, s_e, is calculated using the following equation:

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2}}$$

DEFINITION

The **standard error of the regression,** s_e**,** measures the amount of dispersion of the observed data around the estimated regression line.

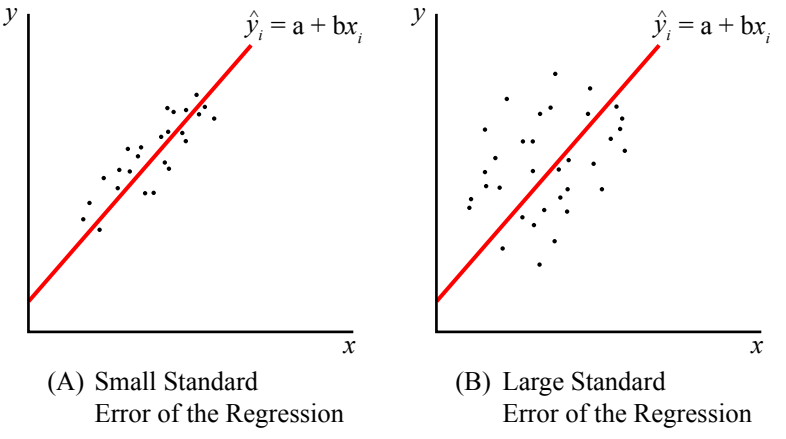

Figure 3.8
Standard error of the regression.

For our home prices example:

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{11{,}934.88}{10-2}} = \sqrt{1{,}491.86} = \$38.62$$

This number means that in about two-thirds of the time (68.3 percent according to the empirical rule. Remember that?) home prices will vary by $38,620 ($38.62 × 1,000) around the estimated prices given in the equation you estimated. You are getting really good here!

Confidence Interval for the Mean Value of Y

One way we can use the estimated regression equation above is to forecast the price of a house based on its size, as we did above. For a 1200 square feet house, we found that the expected average price is $219,873. However, we know that different 1200 square feet homes will have different prices due to other factors that are not included in our regression line, such as the age of the house, the finishing, the included appliances, and many other things. So just how accurate is our predicted price for a given house size? To answer that question, we will estimate a confidence interval (sounds familiar?) around the estimated y value for a given value of x. This confidence interval will give us a range within which we are certain that the average price will be for all 1200 square feet houses.

In general, the confidence interval around the mean of y given a specific value of x can be found by:

$$CI = \hat{y} \pm t_c s_e \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} x_i^2\right) - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}}$$

where:

t_c = the critical t-value from the Students' t-distribution

s_e = the standard error of the regression

n = the number of observations

Hold on to your hat while we dive into this one with our example. Suppose we would like a 95 percent confidence interval around the mean of y for $x = 1200$ square feet. To find our critical t-value, we look to Table 4 in Appendix B. This procedure has $n - 2 = 10 - 2 = 8$ degrees of freedom, resulting in $t_c = 2.306$ from Table 4 in Appendix B. Our confidence interval is then:

$$CI = \hat{y} \pm t_c s_e \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} x_i^2\right) - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}}$$

$$CI = 219.873 \pm (2.306)(38.62)\sqrt{\frac{1}{10} + \frac{(1{,}200 - 1{,}676.5)^2}{(30{,}860{,}525) - \frac{(16{,}765)^2}{10}}}$$

$CI = 219.873 \pm (2.306)(38.62)(0.427) = 219.873 \pm 38.028$

$CI = 181.845$ and 257.901

This interval is shown graphically in Figure 3.9.

Our 95 percent confidence interval for the average 1200 square feet home price is between \$181.845 and \$257.901 thousand. This means that we are 95 percent confident that the average price of all 1200 square feet homes in this neighborhood is between \$181,845 and \$257,901.

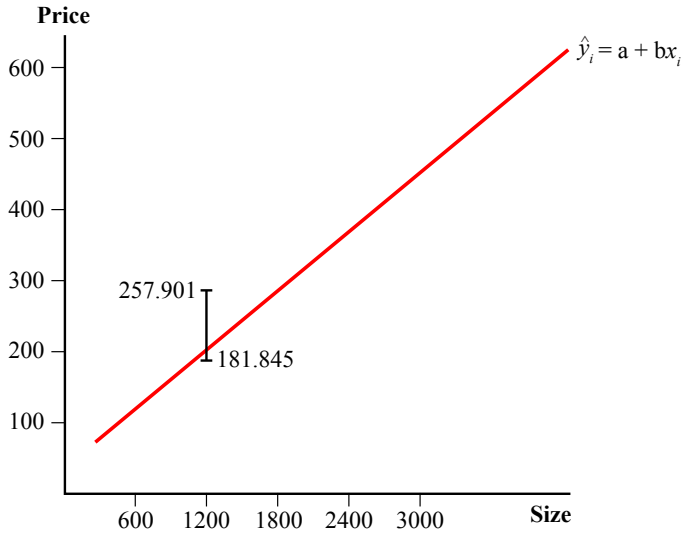

Figure 3.9
95 percent confidence interval for the average y given x = 1200.

Hypothesis Testing and the Confidence Interval for the Coefficient of the Regression Line

Recall that if the slope of the regression line, *b*, is equal to 0, then there is no relationship between *x* and *y*. In our home prices example, we found the slope of the regression line to be 0.15914. However, because this result was based on a sample of observations, we need to test whether 0.15914 is far enough away from 0 to claim that a relationship really does exist between the variables. If β_1 is the slope of the true population, then our hypotheses statement would be:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

If we reject the null hypothesis, we conclude that a relationship does exist between the independent and dependent variables based on our sample. We'll test this using $\alpha = 0.01$.

This hypothesis test requires the standard error of the slope, s_b, which is found with the following equation:

$$s_b = \frac{s_e}{\sqrt{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}}$$

where s_e is the standard error of the regression that we calculated earlier.

For our home prices example:

$$s_b = \frac{s_e}{\sqrt{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}} = \frac{38.62}{\sqrt{30{,}860{,}525 - 10(1{,}676.5)^2}} = \frac{38.62}{\sqrt{2{,}754{,}002}} = 0.02327$$

The calculated t-test statistic for this hypothesis is:

$$t = \frac{b - B_1}{s_b}$$

where B_1 is the value of the population slope according to the null hypothesis.

For this example, our calculated t-test statistic is:

$$t = \frac{b - B_1}{s_b} = \frac{0.15914 - 0}{0.02327} = 6.8388$$

The critical t-value is taken from the Student's t-distribution with $n - 2 = 10 - 2 = 8$ degrees of freedom. With a two-tail test and $\alpha = 0.01$, $t_c = 3.355$ according to Table 4 in Appendix B. Because the calculated t-test statistic $t >$ the critical t-value t_c, we reject the null hypothesis and conclude that there is a relationship between the size of the house and its price.

TEST YOUR KNOWLEDGE

I know you might be wondering "Why are we using $n - 2$ as the degrees of freedom here?" The degrees of freedom = n – the number of estimated coefficients. Since we are estimating two coefficients here, the slope b and the y-intercept a, then the degrees of freedom = $n - 2$.

I hear you asking "Can we get a confidence interval for the actual population slope (B_1) using the sample slope b?" Yes, just like any confidence interval, we use the following equation:

$$CI = b \pm t_{\alpha/2}\, s_b$$

So let's apply it to our example. The value for $t_{\alpha/2}$ is taken from the Student's t-distribution with $n - 2 = 10 - 2 = 8$ degrees of freedom. With a 95 percent confidence level, $t_{\alpha/2} = 2.306$ according to Table 4 in Appendix B. The 95 percent CI is

$$CI = b \pm t_{\alpha/2}\, s_b = 0.15914 \pm (2.306)(0.02327)$$

$$CI = 0.15914 \pm 0.053661 = (0.105479, 0.212801)$$

This means that we are 95 percent confident that the price of a house will increase by between \$106 and \$213 for every one square foot increase in its size. How useful is this information for you and your spouse when buying a house? Now you can prove to your spouse that you know statistics pretty well! After all, you have read the entire book, haven't you?

Using Excel to Perform Simple Linear Regression

Now that we have burned out our calculators with all these fancy equations, let me show you how Excel does it all for us. You will be surprised by how easy it is this way!

1. Start by placing the data for home prices and sizes in Columns A and B in a blank sheet.
2. On the Tools menu at the top of Excel window, click on the Data tab and select Data Analysis. (Refer to the section "Installing the Data Analysis Add-In" from Chapter 2 if you don't see the Data Analysis command on the Tools menu.)
3. From the Data Analysis dialog box, select Regression as shown in Figure 3.10 and click OK.

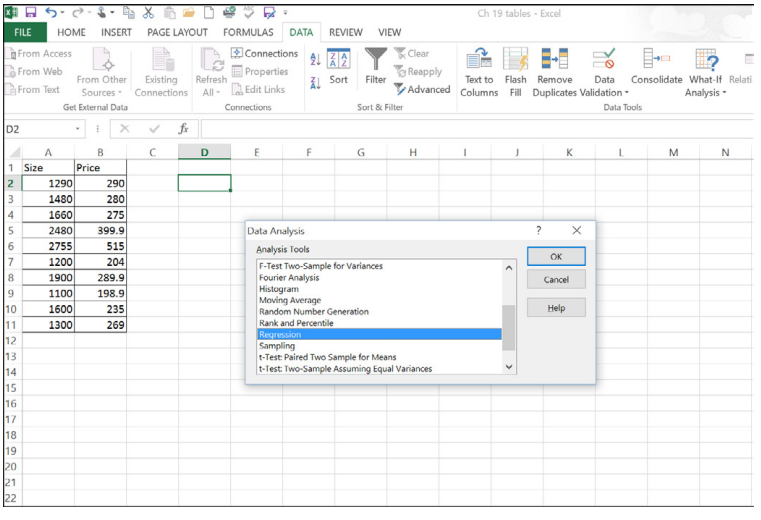

Figure 3.10
Setting up simple linear regression with Excel.

4. Set up the Regression dialog box according to Figure 3.11.

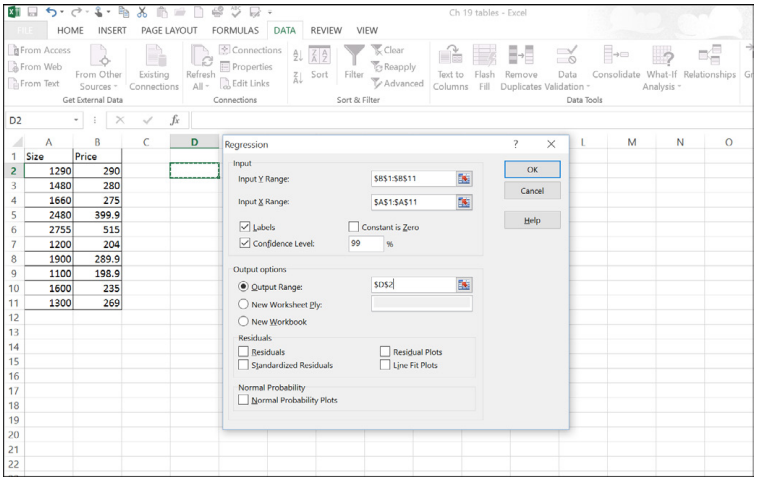

Figure 3.11

The Regression dialog box for the home prices example.

5. Click OK. Figure 3.12 shows the final regression results.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.924054982
R Square	0.85387761
Adjusted R Square	0.835612312
Standard Error	38.66961009
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	69905.03104	69905.031	46.74863	0.000132689
Residual	8	11962.70996	1495.3387		
Total	9	81867.741			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	28.56894148	40.93448049	0.6979188	0.505004	-65.8261398	122.9640228	-108.782096	165.9199787
Size	0.159320643	0.023301701	6.8372967	0.000133	0.105586824	0.213054463	0.08113441	0.237506876

Figure 3.12

The final results of regression analysis in Excel for the home prices example.

These results are consistent with what we found after grinding it out in the previous sections. Because the *p*-value for the independent variable size is shown as 0.000133, which is less than $\alpha = 0.01$, we can reject the null hypothesis and conclude that a relationship between the variables does exist.

Now, let's look at Excel regression output in detail and see what each number means:

- The Regression Statistics table: Multiple *R* is the correlation coefficient between *x* and *y*; *R* Square is the coefficient of determination R^2; Standard Error is the standard error of the regression *se* (R^2 and s_e are the measures of goodness-of-fit for the model); and Observations is the number of observations.
- The ANOVA table: the SS column presents the sum of squares; the df column presents the degrees of freedom; under the SS column and across from Regression is the explained variation SSR; under the SS column and across from Residual is the unexplained variation SSE; and under the SS column and across from Total is the total Variation, SST. The MS column presents the mean squares, which is the sum of squares, SS, divided by degrees of freedom, d.f.
- The last table: presents the estimated coefficients; their standard errors; their *t*-test statistics; their *p*-values; and their confidence intervals. For the slope, for example, the estimated slope (*b*) is 0.15932 (the small difference between this number and what we calculated above is due to rounding); its standard error, s_b, is 0.0233; the *t*-test statistic is 6.837; its *p*-value is 0.000133; and the 95 percent confidence interval for the slope is (0.106, 0.213).

Note that in the Excel Regression dialog box, I marked the box for confidence level and chose 99 percent (shown in Figure 3.11). By doing this, I get two confidence intervals: the 95 percent (which comes automatically with Excel regression output) and the 99 percent (that I chose). The 99 percent confidence interval for the slope is (0.081, 0.238).

I hear you asking about the three numbers that I skipped: Adjusted *R* Square, *F*, and Significance *F*. Those are relevant to the multiple regression, which is outside the scope of this book. However, for the simple linear regression, the Significance *F* is equivalent to the *p*-value for the slope. Looking at Figure 3.12, you can confirm that they are both equal to 0.000133.

Now we know you must be very excited about learning how to use Excel to perform a regression analysis and how to interpret the results, so we'll give you another example with a negative correlation.

A Simple Linear Regression Example with a Negative Correlation

Both of these past examples have involved a positive relationship between *x* and *y*. Now this example will summarize performing simple linear regression with a negative relationship.

Bob had the opportunity to "bond" with his son Brian when he was buying him his first car when Brian turned 16. Brian had visions of Mercedes and BMWs dancing in his head, whereas Bob was thinking more along the line of Hondas and Toyotas. After many "discussions" on the matter,

they agreed on looking for Volkswagen Jettas. The following table shows the mileage of eight cars and their asking price. The remainder of this chapter demonstrates the correlation and regression technique using this data.

Data for Car Example

Mileage	Price	Mileage	Price
21,800	$16,000	65,800	$10,500
34,000	$11,500	72,100	$12,300
41,700	$13,400	76,500	$8,200
53,500	$14,800	84,700	$9,500

The following table, which shows the data in thousands, will be used for the various equations.

Mileage (x)	Price (y)	xy	x^2	y^2
21.8	16.0	348.80	475.24	256.00
34.0	11.5	391.00	1156.00	132.25
41.7	13.4	558.78	1738.89	179.56
53.5	14.8	791.80	2862.25	219.04
65.8	10.5	690.90	4329.64	110.25
72.1	12.3	886.83	5198.41	151.29
76.5	8.2	627.30	5852.25	67.24
84.7	9.5	804.65	7174.09	90.25
$\sum x = 450.1$	$\sum y = 96.2$	$\sum xy = 5{,}100.1$	$\sum x^2 = 28{,}786.8$	$\sum y^2 = 1{,}205.9$

$$\bar{x} = \frac{450.1}{8} = 56.3 \qquad \bar{y} = \frac{96.2}{8} = 12.0$$

The correlation coefficient can be found using:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n(\Sigma x^2) - (\Sigma x)^2\right]\left[n(\Sigma y^2) - (\Sigma y)^2\right]}}$$

$$r = \frac{8(5{,}100.1) - (450.1)(96.2)}{\sqrt{\left[8(28{,}786.8) - (450)^2\right]\left[8(1{,}205.9) - (96.2)^2\right]}} = \frac{-2{,}498.82}{\sqrt{(27{,}794.4)(392.76)}}$$

$r = -0.756$

The negative correlation indicates that as mileage (x) increases, the price (y) decreases as we would expect. The coefficient of determination R^2 can also be calculated as the square of the correlation coefficient (r), so it becomes:

$$R^2 = (r)^2 = (-0.756)^2 = 0.572$$

Approximately 57 percent of the variation in price is explained by the variation in mileage. The regression line is determined using:

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{8(5{,}100.1) - (450.1)(96.2)}{8(28{,}786.8) - (450.1)^2} = \frac{-2{,}498.82}{27{,}704.39}$$

$$b = -0.0902$$

$$a = \overline{y} - b\overline{x} = 12.025 - (-0.0902)(56.26) = 17.100$$

We can describe the estimated regression line by the equation:

$$\hat{y}_i = 17.1 - 0.0902\, x_i$$

This equation is shown graphically in Figure 3.13.

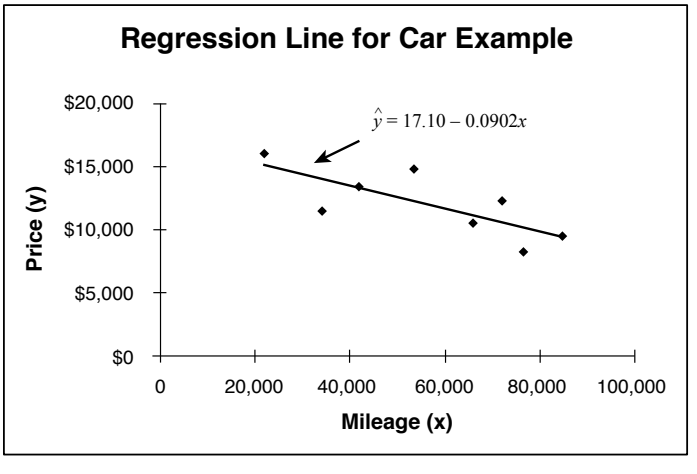

Figure 3.13
Regression line for the car example.

What would the predicted price be for a car with 45,000 miles?

$$\hat{y} = 17.1 - 0.0902(45.0) = \$13{,}041$$

The regression line would predict that a car with 45,000 miles would be priced at $13,041. What would be the 90 percent confidence interval at $x = 45{,}000$? The standard error of the regression would be:

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2}}$$

$$s_e = \sqrt{\frac{(1{,}205.9) - (17.1)(96.2) - (-0.0902)(5{,}100.1)}{8-2}} = 1.867$$

The critical t-statistic for $n - 2 = 8 - 2 = 6$ degrees of freedom and a 90 percent confidence interval is $t_c = 1.943$ from Table 4 in Appendix B. Our confidence interval is then:

$$CI = \hat{y} \pm t_c s_e \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} x_i^2\right) - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}}$$

$$CI = 13.041 \pm (1.934)(1.867)\sqrt{\frac{1}{8} + \frac{(45 - 56.26)^2}{(28{,}786.8) - \frac{(450.1)^2}{8}}}$$

$CI = 13.041 \pm (1.934)(1.867)(0.402) = 13.041 \pm 1.452$

$CI = 11.589$ and 14.493

The 90 percent confidence interval for the average price of cars with 45,000 miles is $11.589 and $14.493 thousand. This means that we are 90 percent confident that the average price of cars with 45,000 miles is between $11,589 and $14,493.

Is the relationship between mileage and price statistically significant at the $\alpha = 0.10$ level? Our hypotheses' statement is:

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

The standard error of the slope, s_b, is found using:

$$s_b = \frac{s_e}{\sqrt{\sum x_i^2 - n(\bar{x})^2}} = \frac{1.867}{\sqrt{28{,}786.8 - 8(56.26)^2}} = 0.0317$$

The calculated t-test statistic for this hypothesis is:

$$t = \frac{b - B_1}{s_b} = \frac{-0.0902 - 0}{0.0317} = -2.845$$

The critical t-value is taken from the Student's t-distribution with $n - 2 = 8 - 2 = 6$ degrees of freedom. With a two-tail test and $\alpha = 0.10$ level, $t_c = \pm 1.943$ according to Table 4 in Appendix B. Because $|t| > |t_c|$, we reject the null hypothesis and conclude there is a relationship between the mileage of the car and its price.

We can also get the 90 percent confidence interval for the slope as:

$$CI = b \pm t_{\alpha/2}\, s_b = -0.0902 \pm (1.943)(0.0317)$$

$$CI = -0.0902 \pm 0.0616 = (-0.1518, -0.0286)$$

This means that we are 90 percent confident that the price of the car will decrease by between \$28.60 and \$151.80 as its mileage increases by 1000 miles.

I know you can't wait to do the analysis with Excel. We will start by placing car mileage and prices data in Columns A and B in a blank sheet. On the menu at the top of the Excel window, click on the Data tab, select Data Analysis, and select Regression. Set up the Regression dialog box according to Figure 3.14.

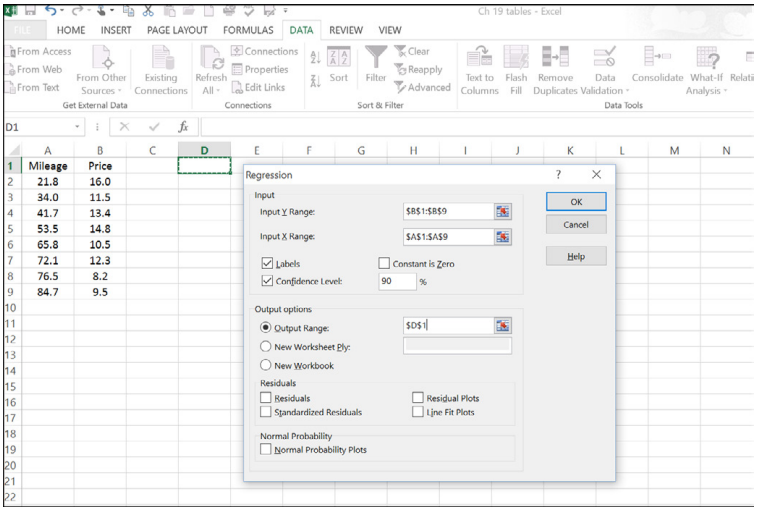

Figure 3.14
The Regression dialog box for the car prices example.

Click OK. Figure 3.15 shows the final regression results.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.757779798
R Square	0.574230222
Adjusted R Square	0.503268593
Standard Error	1.866130213
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	28.18034816	28.18035	8.092123	0.02938649
Residual	6	20.89465184	3.482442		
Total	7	49.075			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	17.1003358	1.902243016	8.989564	0.000106	12.4457148	21.7549568	13.40393468	20.79673692
Mileage	-0.09020815	0.031711335	-2.84467	0.029386	-0.16780299	-0.0126133	-0.15182899	-0.028587305

Figure 3.15

The final results of regression analysis in Excel for the car prices example.

These results are consistent with what we found using the formulas. Because the *p*-value for the independent variable Mileage is shown as 0.0294, which is less than $\alpha = 0.10$, we can reject the null hypothesis and conclude that a relationship between the variables does exist. Excel makes regression analysis easy and fun!

TEST YOUR KNOWLEDGE

Do you wonder why the Multiple *R* value, which is the correlation coefficient in Figure 3.15, is positive and not negative? This is because Multiple *R* provided in Excel regression output is actually the correlation coefficient between *y* and $\hat{y}$. For a simple linear regression, it is the equivalent to the correlation coefficient between the absolute values of *x* and *y*. In other words, *r* between *y* and $\hat{y}$ is 0.75778, whereas *r* between *y* and *x* is -0.75778.

Assumptions for Simple Linear Regression

For all these results to be valid, we need to make sure that the underlying assumptions of the simple linear regression are not violated. These assumptions are as follows:

- Individual differences between the data and the regression line, $(y_i - \hat{y}_i)$, are independent of one another.
- The residuals are normally distributed with a mean of zero.
- The variation of *y* around the regression line is equal for all values of *x*. In other words, the residuals have a constant variance.
- The independent variable(s) is not correlated with the error term.

- The observed values of y are normally distributed around the predicted value, $\hat{y}$.

Unfortunately (or fortunately), the techniques to test these assumptions go beyond the level of this book.

Simple vs. Multiple Regression

Simple regression is limited to examining the relationship between a dependent variable and only one independent variable. If more than one independent variable is involved in the relationship, then we need to graduate to multiple regression. The regression equation for this method looks like this:

$$\hat{y} = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

As you can imagine, the calculation for this technique gets *really* messy and goes beyond the scope of this book.

This concludes our journey through statistics. We hope you enjoyed it as much as we did!

Practice Problems

1. The following table shows the payroll for 10 major league baseball teams (in millions) for the 2002 season, along with the number of wins for that year.

Payroll	Wins	Payroll	Wins
$171	103	$56	62
$108	75	$62	84
$119	92	$43	78
$43	55	$57	73
$58	56	$75	67

Calculate the correlation coefficient. Test to see whether the correlation coefficient is not equal to 0 at the 0.05 level.

2. Using the data from Problem 1, answer the following questions:
   a) What is the regression line that best fits the data?
   b) Is the relationship between payroll and wins statistically significant at the 0.05 level?
   c) What is the predicted number of wins with a $70 million payroll?
   d) What is the 99 percent confidence interval around the mean number of wins for a $70 million payroll?
   e) What percent of the variation in wins is explained by the payroll?

3. The following table shows the grade point average (GPA) for five students along with their entrance exam scores for MBA programs (GMAT). Develop a model that would predict the GPA of a student based on his GMAT score. What would be the predicted GPA for a student with a GMAT score of 600?

Student	GPA	GMAT
1	3.7	660
2	3.0	580
3	3.2	450
4	4.0	710
5	3.5	550

The Least You Need to Know

- The correlation coefficient, *r,* indicates both the strength and direction of the relationship between the independent and dependent variables.
- The technique of simple linear regression enables us to describe a straight line that best fits the data for two (or more) variables.
- The ordinary least squares (OLS) method is a mathematical procedure to identify the linear equation that best fits the data by finding values for a, the y-intercept; and b, the slope.
- The standard error of the regression, s_e, measures the amount of dispersion of the observed data around the regression line.
- The coefficient of determination, R^2, represents the percentage of the variation in y that is explained by the regression line.