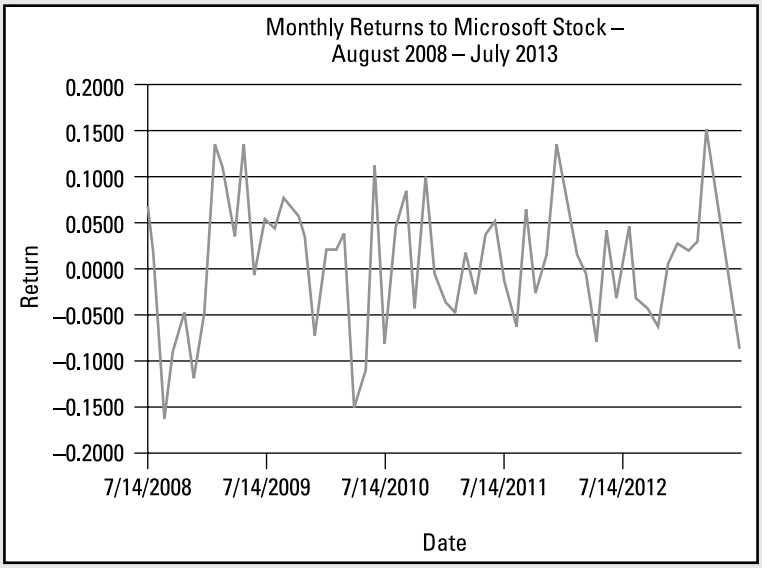
# Part IV

# More Advanced Techniques: Regression Analysis and Forecasting




web extras

Learn about more forecasting and regression analysis at www.dummies.com/extras/businessstatistics.

## In this part...

- Use the powerful technique of regression analysis to estimate the relationship between two variables, and take an in-depth look at multiple regression where a single dependent variable depends on two or more independent variables.
- Understand the effects of seasonal variation on sales of everything from gasoline prices to retail items. Use a scatterplot to see if a time series exhibits seasonal variation and, if so, what type.
- Predict the future values of economic variables, including stock prices, interest rates, and more.

# Chapter 15

# Simple Regression Analysis

### In This Chapter

- Understanding the assumptions underlying regression analysis
- Implementing the simple regression model
- Interpreting the regression results

*Regression analysis* is a statistical methodology that helps you estimate the strength and direction of the relationship between two or more variables. The two types of regression analysis are *simple regression analysis* (which I discuss in this chapter) and *multiple regression analysis* (which I cover in Chapter 16). Simple regression analysis allows you to estimate the relationship between a dependent variable (*Y*) and an independent variable (*X*). Multiple regression analysis allows you to estimate the relationship between a dependent variable (*Y*) and two or more independent variables (*X*s).

For example, suppose a researcher is interested in analyzing the relationship between the annual returns to the Standard & Poor's 500 (S&P 500) and the annual returns to Apple stock.


The Standard and Poor's 500 (S&P 500) is a broad-based stock market index; it contains the 500 largest U.S. stocks, based on *market capitalization*. (The market capitalization of a stock equals the market price of the stock times the number of outstanding shares.) The returns to the S&P 500 are often used to represent the performance of the U.S. stock market.

The researcher assumes that the returns to Apple stock are at least partially explained by the returns to the S&P 500 because the S&P reflects overall activity in the economy. In other words, the researcher assumes that the return on Apple stock depends on the returns to the S&P 500.

To analyze this relationship with simple regression analysis, you treat the returns on Apple stock as a dependent variable (*Y*) and the returns to the S&P 500 as an independent variable (*X*). Regression analysis makes it possible to determine *how much* the returns on Apple stock are affected by the returns to the S&P 500. (In other words, how strong is the relationship between Apple stock and the S&P 500.)

This chapter introduces the basic regression analysis framework, including the underlying assumptions and the formulas you need to estimate the relationships between different variables. I also cover techniques for testing the validity of the results in great detail.

# The Fundamental Assumption: Variables Have a Linear Relationship

Simple regression analysis is based on the assumption that a linear relationship exists between $X$ and $Y$. Intuitively, if two variables have a linear relationship between them, a graph of the two variables is a straight line. (For a more formal discussion of linear relationships, see the following section "Defining a linear relationship.")

For example, suppose that an equity analyst at a prestigious investment bank wants to determine the relationship between a corporation's sales and profits to help him estimate the proper value of the corporation's stock. He has reason to believe that the relationship between sales and profits is linear. Further, he assumes that profits are the dependent variable in this relationship, while sales are the independent variable. Specifically, he believes that each $1,000 increase in sales triggers an increase in profits by $200, while each $1,000 decrease in sales has the opposite effect.

The analyst may use regression analysis to determine the actual relationship between these variables by looking at the corporation's sales and profits over the past several years. The regression results show whether this relationship is valid. In addition to sales, other factors may also determine the corporation's profits, or it may turn out that sales don't explain profits at all. The regression results also show the estimated amount that the profits change when sales change by $1,000.

In the following sections, I dig deeper into the linear relationship between the dependent and independent variables and show you how to represent this relationship graphically.

## Defining a linear relationship

In terms of geometry, you can graph a linear relationship with a straight line. Algebraically, the general expression for a linear relationship is

$$Y = mX + b$$


$X$ is the independent variable, $Y$ is the dependent variable whose value is determined by the value of $X$, $m$ is the slope coefficient (how much $Y$ changes in response to a change in $X$), and $b$ is the intercept (the value of $Y$ if $X$ equals 0).

You calculate the slope of a line ($m$) with this formula:

$$m = \frac{\Delta Y}{\Delta X}$$

Here, $\Delta Y$ ("delta $Y$") represents the change in $Y$, and $\Delta X$ ("delta $X$") represents the change in $X$.

Think of the slope as a measure of how much $Y$ changes due to a given change in $X$, or how *sensitive* the value of $Y$ is to changes in $X$. A linear relationship is one in which the slope is a *constant*.

You see a linear relationship graphed as a straight line, with the dependent variable ($Y$) on the vertical axis and the independent variable ($X$) on the horizontal axis. See Figure 15-1 for the relationship between $X$ and $Y$ in the equation $Y = 2X + 3$.

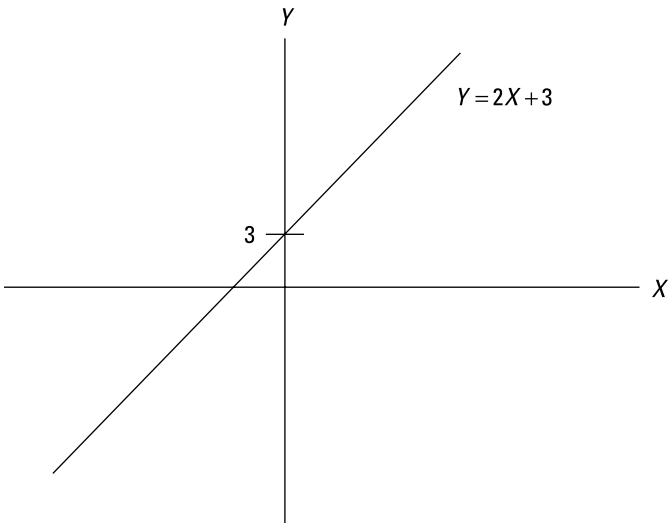



**Figure 15-1:** Graph of a linear relationship: $Y = 2X + 3$.

The equation of the line, $Y = 2X + 3$, tells you two important things:

- The *slope* of the line is 2 (this is the constant that's multiplied by $X$), which shows that
  - For each increase in $X$ by 1, $Y$ increases by 2.
  - For each decrease in $X$ by 1, $Y$ decreases by 2.
- The *intercept* of the line is 3, so if $X = 0$, the value of $Y$ is 3. (In Figure 15-1, you see that 3 is the point where the line crosses the $Y$ axis.)

## Using scatter plots to identify linear relationships

A *scatter plot* is a special type of graph designed to show the relationship between two variables. (See Chapter 5 for an introduction to scatter plots.)

With regression analysis, you can use a scatter plot to visually inspect the data to see whether $X$ and $Y$ are linearly related. The following are some examples.

Figure 15-2 shows a scatter plot for two variables that have a *nonlinear* relationship between them.

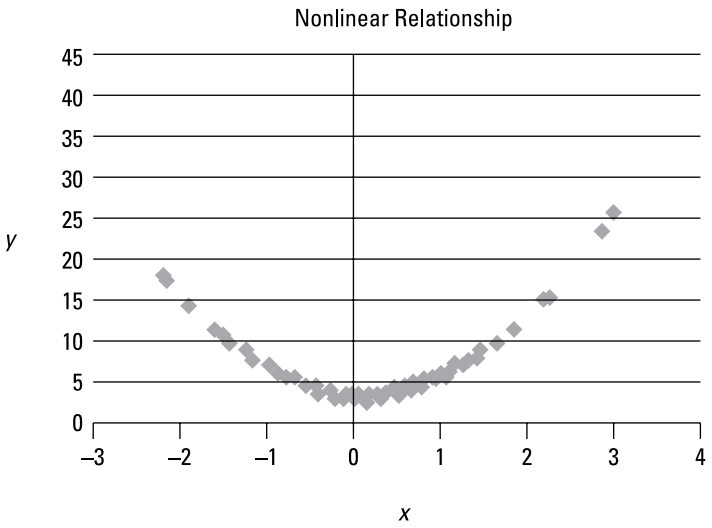



**Figure 15-2:** Scatter plot of a nonlinear relationship.

Each point on the graph represents a single (*X*, *Y*) pair. Because the graph isn't a straight line, the relationship between *X* and *Y* is nonlinear. Notice that starting with negative values of X, as *X* increases, *Y* at first decreases; then as X continues to increase, *Y* increases. The graph clearly shows that the slope is continually changing; it isn't a constant. With a linear relationship, the slope never changes.

In this example, one of the fundamental assumptions of simple regression analysis is violated, and you need another approach to estimate the relationship between *X* and *Y*. One possibility is to transform the variables; for example, you could run a simple regression between ln(X) and ln(Y). ("ln" stands for the natural logarithm.) This often helps eliminate nonlinearities in the relationship between X and Y. Another possibility is to use a more advanced type of regression analysis, which can incorporate nonlinear relationships.


One regression technique that may be used with nonlinear data is known as *nonlinear least squares* (details may be found at `https://en.wikipedia.org/wiki/Non-linear_least_squares`).

Figure 15-3 shows a scatter plot for two variables that have a strongly positive linear relationship between them. The correlation between *X* and *Y* equals 0.9. (See Chapter 5 for an overview on correlation.)

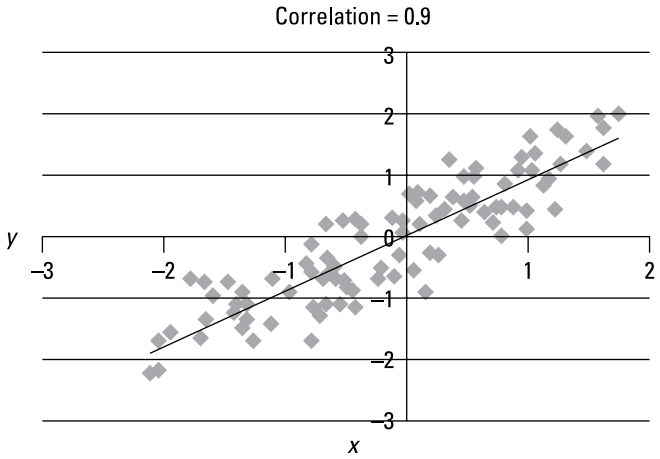



**Figure 15-3:** Scatter plot of a strongly positive linear relationship.

Figure 15-3 shows a very strong tendency for *X* and *Y* to both rise above their means or fall below their means at the same time. The straight line is a *trend line,* designed to come as close as possible to all the data points. The trend

line has a positive slope, which shows a positive relationship between *X* and *Y*. The points in the graph are tightly clustered about the trend line due to the strength of the relationship between X and Y. (***Note:*** The slope of the line is *not* 0.9; 0.9 is the correlation between X and Y.)

Figure 15-4 shows a scatter plot for two variables that have a weakly positive linear relationship between them; the correlation between *X* and *Y* equals 0.2

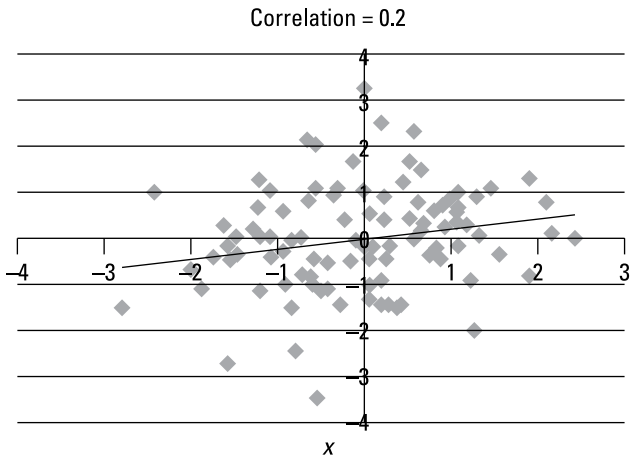



**Figure 15-4:** Scatter plot of a weakly positive linear relationship.

Figure 15-4 shows a weaker connection between *X* and *Y*. Note that the points on the graph are more scattered about the trend line than in Figure 15-3, due to the weaker relationship between X and Y.

Figure 15-5 is a scatter plot for two variables that have a strongly negative linear relationship between them; the correlation between *X* and *Y* equals –0.9.

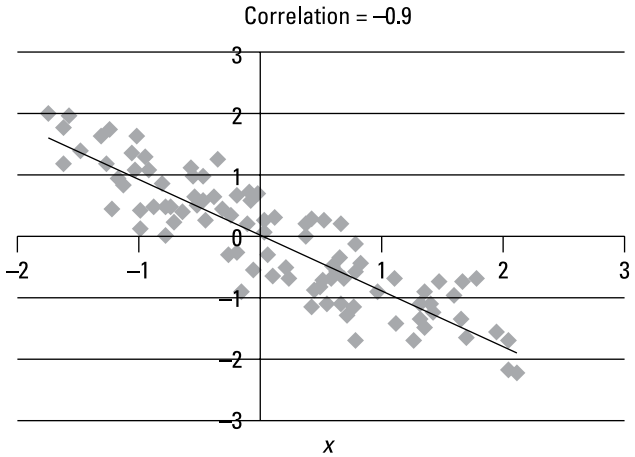



**Figure 15-5:** Scatter plot of a strongly negative linear relationship.

Figure 15-5 shows a very strong tendency for *X* and *Y* to move in opposite directions; for example, rise above or fall below their means at opposite times. The trend line has a negative slope, which shows a negative relationship between *X* and *Y*. The points in the graph are tightly clustered about the trend line due to the strength of the relationship between X and Y.

Figure 15-6 is a scatter plot for two variables that have a weakly negative linear relationship between them. The correlation between *X* and *Y* equals –0.2.

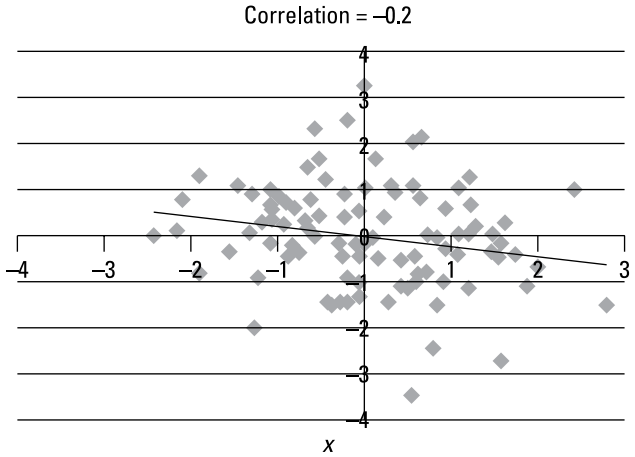



**Figure 15-6:** Scatter plot of a weakly negative linear relationship.

Figure 15-6 shows a very weak connection between *X* and *Y*. Note that the points on the graph are more scattered about the trend line than in Figure 15-5 due to the weaker relationship between X and Y.

# Defining the Population Regression Equation

With regression analysis, you typically draw a sample of data from a population to estimate the relationship between *X* and *Y*. The equation that best explains the population data is known as the *population regression equation,* or *population regression line:*

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$


The symbol β is the Greek letter "beta," and the symbol ε is "epsilon." $\beta_0$ and $\beta_1$ are known as *coefficients* of the regression line. $\beta_1$ is the slope coefficient and $\beta_0$ is the intercept coefficient (or simply the intercept). A coefficient is a constant that is multiplied by a variable.

Based on the assumption that the relationship between X and Y is linear, the regression line is designed to capture this relationship as closely as possible.

Other key terms in the equation are

- $i$ = an index used to identify the members of the population.
- $Y_i$ = a single value of $Y$, indexed by $i$, in a population of size $n$, with the values of $Y$ expressed as $Y_1, Y_2, Y_3, \ldots, Y_n$.
- $X_i$ = a single value of $X$, indexed by $i$, in a population of size $n$, with the values of $X$ expressed as $X_1, X_2, X_3, \ldots, X_n$.
- $\varepsilon_i$ = an "error term," indexed by $i$; each observation in the population $(X_i, Y_i)$ has an error term associated with it.

Using the example of the equity analyst from the earlier section, "The Fundamental Assumption: Variables Have a Linear Relationship," suppose that the corporation has been in business for the past ten years (2003 to 2012). $X_1$ represents sales in 2003, and $Y_1$ represents profits in 2003. $X_2$ represents sales in 2004, and $Y_2$ represents profits in 2004. The analyst continues through 2012, where $X_{10}$ is 2012 sales, and $Y_{10}$ is 2012 profits. Each $(X_i, Y_i)$ pair is a single observation chosen from the population.

The population regression equation has a slope and an intercept and one other term that you don't normally find in the equation for a straight line — the *error term.* The error term is included because the population regression equation doesn't perfectly capture the relationship between $X$ and $Y$. For example, suppose that in the population regression line, $\beta_0 = 10$ and $\beta_1 = 2$. Assume that actual year 2003 sales were \$100 million. The population regression line indicates that profits in 2003 should be

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1(X_1) \\ Y_1 &= 10 + (2)(100) \\ &= \$210 \text{ million} \end{aligned}$$

Suppose that 2003 profits were actually \$200 million. The population regression line *overstates* actual 2003 sales by \$10 million. As a result, you compute the error term for 2003 ($\varepsilon_1$) as follows:

$$\begin{aligned} Y_1 &= 10 + 2X_1 + \varepsilon_1 \\ \varepsilon_1 &= Y_1 - 10 - 2X_1 \\ \varepsilon_1 &= 200 - 10 - 2(100) \\ \varepsilon_1 &= 200 - 10 - 200 \\ \varepsilon_1 &= -10 \end{aligned}$$

# Estimating the Population Regression Equation

In most situations, estimating the population regression line with the entire population is impractical because collecting the amount of required data can be expensive and time-consuming. Instead, you draw a sample from the underlying population that reflects the underlying population as closely as possible). You use the sample data to construct a *sample regression equation,* or *sample regression line,* which you then use as an estimate of the actual population regression equation. (Sampling techniques and sampling distributions are discussed in Chapter 10.)

The sample regression equation is expressed as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Here, $\hat{Y}_i$ is the estimated value of $Y_i$, associated with $X_i$ $\hat{\beta}_0$ is the estimated value of $\beta_0$, and $\hat{\beta}_1$ is the estimated value of $\beta_1$.


Note that there is no estimated error term in this equation because the estimated value of $Y_i$ is actually the average value of a probability distribution; thus, there is no error term associated with it.


The symbol ^ often indicates an *estimated value.* The proper name for this punctuation mark is *caret.* Often, it's informally called a "hat." For example, you pronounce $\hat{\beta}_0$ as "beta zero hat."

You determine these estimated values for $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing the sum of the squared differences between the actually observed *Y* values contained in the sample data and those that have been *predicted* by the sample regression equation, as shown in the following equation:

$$\min \sum_{i=1}^{n} \left( Y_i - \hat{Y}_i \right)^2$$

***Note:*** In this formula, *min* stands for "minimize" and tells you to choose values of $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the predicted values of Y are as close as possible to the actual values of Y. Think of each term

$$\left( Y_i - \hat{Y}_i \right)$$

as a potential mistake or error by the regression line. If this term is *positive,* the regression line has *underestimated* the true value of $Y_i$. If this term is

*negative*, the regression line has *underestimated* the true value of $Y_i$. If this term equals zero, the regression line has correctly estimated the true value of $Y_i$.

The objective of regression analysis is to find the equation that minimizes the sum of these errors.


Note that the value being minimized is actually the sum of the *squared* values of $\left(Y_i - \hat{Y}_i\right)$. This is because the sum of these terms always equals zero.

The difference between the actual value of $Y_i$ and the predicted value of $Y_i$ is known as a *residual* — an estimate of the corresponding error term in the population regression equation — and is expressed as follows:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

$\hat{\varepsilon}_i$ represents the residual associated with a single observation from the population ($X_i$, $Y_i$).

As an example, suppose the quality control manager for a manufacturing company is interested in seeing the relationship between annual costs of production and total output for a specific product. She estimates a regression equation based on production data for the years 2005 to 2012. In this case, $X_i$ represents quantity produced during a given year, and $Y_i$ represents total costs during the same year. X represents the quantity produced and Y represents the total costs because costs depend on output, not the other way around.

The manager assigns indexes to the years in the sample as follows: 2005 = Year 1, 2006 = Year 2, 2007 = Year 3, and so forth.

Based on the production data taken from the years 2005 to 2012, the estimated regression equation is

$$\hat{Y}_i = 3 - 1.5X_i$$

The diagram in Figure 15-7 shows the relationship between the actual value of *Y,* the predicted value of *Y,* the mean of *Y,* and the residual for Year 1 (2005).

The variables in this diagram are:

$X_1$ is total output during Year 1.

$Y_1$ is total cost during Year 1.

$\hat{Y}_1$ is the estimated total cost during Year 1.

$\overline{Y}$ is known as "Y bar" and is the average value of *Y* during the sample period.

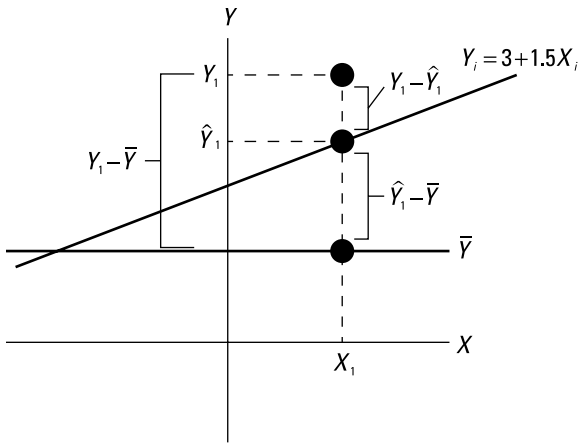



**Figure 15-7:** Predicted value of *Y* versus actual value of *Y*.

Notice that the actual value of $Y_1$ is greater than the value estimated by the regression line. Both values are greater than the average or mean value of Y. (This information is used to construct a measure that explains how well the regression line matches the sample data in the later section "Computing the coefficient of determination.")

For each year's production data from 2005 to 2012,

- $Y_i - \hat{Y}_i$ is the difference between the actual and estimated total cost in Year *i*.
- $\hat{Y}_i - \overline{Y}$ is the difference between the estimated total cost in Year *i* and the average total cost during the sample period.
- $Y_i - \overline{Y}$ is the difference between the total cost in Year *i* and the average total cost during the sample period.

Note that $Y_i - \overline{Y} = \left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \overline{Y}\right)$.

- $\left(Y_i - \hat{Y}_i\right)$ is the size of the incorrect prediction (error) by the regression equation. It equals the difference between the actual value of *Y* and the value predicted by the regression equation.
- $\left(\hat{Y}_i - \overline{Y}\right)$ shows the benefit of using this regression equation to predict the value of $Y_i$ instead of using an alternative, such as simply assuming that each value of $Y_i$ equals $\overline{Y}$.

You estimate the regression equation with formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of the squared residuals:

$$\min \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

The resulting equations for the slope of the estimated regression equation is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

And the equation for the intercept is

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$


These formulas are known as *ordinary least squares* (OLS) estimators. OLS is a methodology for estimating regression coefficients. Some of the more advanced versions include generalized least squares (GLS) and weighted least squares (WLS).


$\overline{X}$ is the mean or average value of $X$; $\overline{Y}$ is the mean or average value of $Y$.

As an example, suppose that $X$ represents the monthly number of hours of studying by college students, and $Y$ represents their corresponding grade point averages (GPAs). To conduct this study, you choose a sample of eight students and list their study hours and GPAs like so:

| ***Y (GPA)*** | ***X (Monthly Hours of Studying)*** |
|---|---|
| 3.5 | 16 |
| 3.2 | 14 |
| 3.0 | 12 |
| 2.6 | 11 |
| 2.9 | 12 |
| 3.3 | 15 |
| 2.7 | 13 |
| 2.8 | 11 |

Then you can create a scatter plot like Figure 15-8 to represent the data.

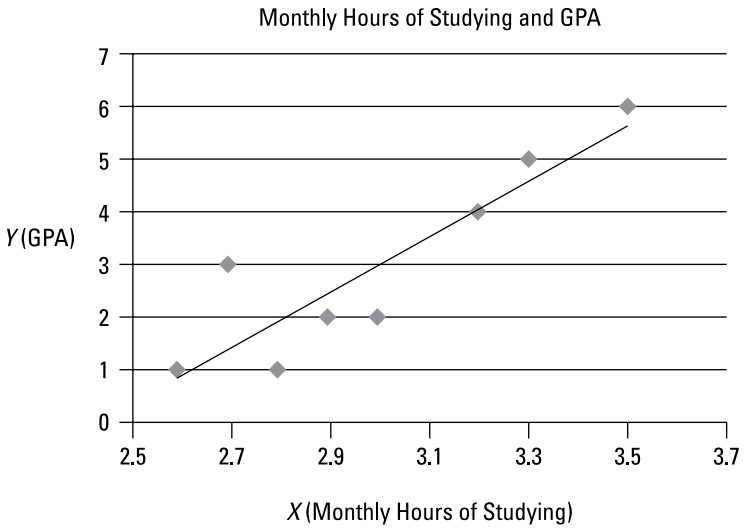



**Figure 15-8:** Scatter plot of monthly study hours and GPA.

Figure 15-8 shows that the relationship between these two variables is approximately linear. As a result, you can estimate the relationship between these two variables with simple regression analysis.

You compute the coefficients of the sample regression equation by following these steps:

1. Find the sample mean of $X$ and $Y$:

$$\begin{aligned}\overline{X} &= \frac{\sum_{i=1}^{n} X_i}{n} \\ &= \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8}{n} \\ &= \frac{16+14+12+11+12+15+13+11}{8} \\ &= 13.0\end{aligned}$$

In this case, you add up the monthly hours of studying for the eight students in the sample and then divide by 8. This gives a sample mean of 13.0 hours for these students.

$$\begin{aligned}\overline{Y} &= \frac{\sum_{i=1}^{n} Y_i}{n} \\ &= \frac{Y_1 + Y_2 + Y_3 + Y_4 + Y_5 + Y_6 + Y_7 + Y_8}{n} \\ &= \frac{3.5 + 3.2 + 3.0 + 2.6 + 2.9 + 3.3 + 2.7 + 2.8}{8} \\ &= 3.0\end{aligned}$$

In this case, you add up the GPAs for the eight students in the sample and then divide by 8. This gives a sample mean of 3.0 for these students.

**The results of the remaining steps are summarized in Table 15-1.**

**Table 15-1 Computing the Regression Slope and Intercept**

| *Y (GPA)* | *X (Monthly Hours of Studying)* | $(X_i - \overline{X})$ | $(X_i - \overline{X})^2$ | $(Y_i - \overline{Y})$ | $(X_i - \overline{X})(Y_i - \overline{Y})$ |
|---|---|---|---|---|---|
| 3.5 | 16 | 3 | 9 | 0.5 | 1.5 |
| 3.2 | 14 | 1 | 1 | 0.2 | 0.2 |
| 3.0 | 12 | –1 | 1 | 0.0 | 0.0 |
| 2.6 | 11 | –2 | 4 | –0.4 | 0.8 |
| 2.9 | 12 | –1 | 1 | –0.1 | 0.1 |
| 3.3 | 15 | 2 | 4 | 0.3 | 0.6 |
| 2.7 | 13 | 0 | 0 | –0.3 | 0.0 |
| 2.8 | 11 | –2 | 4 | –0.2 | 0.4 |
| **Sum** | | | **24** | | **3.6** |

2. To compute $(X_i - \overline{X})$, you subtract the mean of *X* from each value of *X*.
3. To find the value of $(X_i - \overline{X})^2$, you square the value of $(X_i - \overline{X})$ for each result you found in the previous step.
4. You calculate $(Y_i - \overline{Y})$ by subtracting the mean of *Y* from each value of Y.
5. You compute $(X_i - \overline{X})(Y_i - \overline{Y})$ multiplying the results in Steps 2 and 4.

The sum in the $(X_i - \overline{X})^2$ column shows that $\sum_{i=1}^{n}(X_i - \overline{X})^2 = 24$. The sum in the $(X_i - \overline{X})(Y_i - \overline{Y})$ column shows that $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = 3.6$.

6. Based on these results, you compute the values of the regression coefficients as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{3.6}{24} = 0.15$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = 3 - (0.15)(13) = 1.05$$

7. You write the estimated (sample) regression equation as

$$\hat{Y}_i = 1.05 + 0.15X_i$$

The slope of this equation shows that for students who study between 11 and 16 hours per month, each additional monthly hour of studying adds 0.15 points to a student's GPA. The intercept may be interpreted to mean that a student who doesn't study at all will have a GPA of 1.05.

You can use the sample regression equation to estimate the GPA that results from a specified number of hours of studying. For example, if a student studies for 15 hours a month, the sample regression equation predicts a GPA of $\hat{Y}_i = 1.05 + 0.15X_i = 1.05 + (0.15)(15) = 3.30$.


When using a regression line to predict the value of Y for a given value of X, don't use any values of X that aren't contained in the sample data. In this example, the regression line is based on values of X between 11 and 16; the results of using a value of X outside of this range is subject to a great deal of uncertainty.

# Testing the Estimated Regression Equation

After you estimate the regression line (see the earlier section "Estimating the Population Regression Equation"), you can do several tests to check the validity of the results. It may be the case that there is no real relationship between the dependent and independent variables; simple regression generates results even if this is the case. It is, therefore, important to subject the regression results to some key tests that enable you to determine if the results are reliable.

In the following sections, I introduce a statistic that is designed to check whether a regression equation makes sense. This is known as the *coefficient of determination*, also known as $R^2$ (R squared). This is used as a measure of

how well the regression equation actually describes the relationship between the dependent variable (Y) and the independent variable (X).

The next technique that may be used to check regression results is a hypothesis test of the coefficients of the regression equation. The steps used to carry out this hypothesis test are similar to those found in Chapter 12, where hypothesis testing is introduced for the first time. This hypothesis test is sometimes known as the "t-test" because the test statistic and critical values are derived from the Student's t-distribution (discussed in Chapter 11). In this case, if the null hypothesis fails to be rejected, this calls into question the validity of the regression results.

## Using the coefficient of determination ($R^2$)

The coefficient of determination, also known as $R^2$, is a statistical measure that shows the proportion of *variation* explained by the estimated regression line. *Variation* refers to the sum of the squared differences between the values of *Y* and the mean value of *Y,* expressed mathematically as

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$$

$R^2$ always takes on a value between 0 and 1. The closer $R^2$ is to 1, the better the estimated regression equation fits or explains the relationship between *X* and *Y*.

The expression $\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$ is also known as the *total sum of squares* (TSS).

This sum can be divided into the following two categories:

- **Explained sum of squares (ESS):** Also known as the *explained variation,* the ESS is the portion of total variation that measures how well the regression equation explains the relationship between *X* and *Y*.

  You compute the ESS with the formula

  $$ESS=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$$

- **Residual sum of squares (RSS):** This expression is also known as *unexplained variation* and is the portion of total variation that measures discrepancies (errors) between the actual values of *Y* and those estimated by the regression equation.

You compute the RSS with the formula

$$RSS = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

The smaller the value of RSS relative to ESS, the better the regression line fits or explains the relationship between the dependent and independent variable.

- **Total sum of squares (TSS):**

  The sum of RSS and ESS equals TSS.

  $$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

  $R^2$ (the coefficient of determination) is the ratio of explained sum of squares (ESS) to total sum of squares (TSS):

  $$R^2 = \frac{ESS}{TSS}$$

  You can also use this formula:

  $$R^2 = 1 - \frac{RSS}{TSS}$$

  Based on the definition of $R^2$, its value can never be negative. Also, $R^2$ can't be greater than 1, so $0 \le R^2 \le 1$.


With simple regression analysis, $R^2$ equals the square of the correlation between $X$ and $Y$.

## Computing the coefficient of determination

The coefficient of determination is used as a measure of how well a regression line explains the relationship between a dependent variable (Y) and an independent variable (X). The closer the coefficient of determination is to 1, the more closely the regression line fits the sample data.

The coefficient of determination is computed from the sums of squares determined in the earlier section "Using the coefficient of determination ($R^2$)." These calculations are summarized in Table 15-2.

**Table 15-2 Computing the Coefficient of Determination ($R^2$)**

| *Yi* | *Xi* | | | | |
|---|---|---|---|---|---|
| 3.5 | 16 | 3.45 | 0.0025 | 0.2025 | 0.25 |
| 3.2 | 14 | 3.15 | 0.0025 | 0.0225 | 0.04 |
| 3.0 | 12 | 2.85 | 0.0225 | 0.0225 | 0.00 |
| 2.6 | 11 | 2.70 | 0.0100 | 0.0900 | 0.16 |
| 2.9 | 12 | 2.85 | 0.0025 | 0.0225 | 0.01 |
| 3.3 | 15 | 3.30 | 0.0000 | 0.0900 | 0.09 |
| 2.7 | 13 | 3.00 | 0.0900 | 0.0000 | 0.09 |
| 2.8 | 11 | 2.70 | 0.0100 | 0.0900 | 0.04 |
| **Sum** | | | **0.1400** | **0.5400** | **0.68** |

To compute ESS, you subtract the mean value of Y from each of the estimated values of Y; each term is squared and then added together:

$$ESS = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 = 0.54$$

To compute RSS, you subtract the estimated value of Y from each of the actual values of Y; each term is squared and then added together:

$$RSS = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = 0.14$$

To compute TSS, you subtract the mean value of Y from each of the actual values of Y; each term is squared and then added together:

$$TSS = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = 0.68$$

Alternatively, you can simply add ESS and RSS to obtain TSS:

$$TSS = ESS + RSS = 0.54 + 0.14 = 0.68$$

The coefficient of determination ($R^2$) is the ratio of ESS to TSS:

$$R^2 = \frac{ESS}{TSS} = \frac{0.54}{0.68} = 0.7941$$

This shows that 79.41 percent of the variation in *Y* is explained by variation in *X*. Because the coefficient of determination can't exceed 100 percent, a value of 79.41 indicates that the regression line closely matches the actual sample data.

# The t-test

Another important test of the results of regression analysis is to determine whether the slope coefficient ($\beta_1$) is different from 0. If the slope coefficient is close to 0, *X* provides little or no explanatory power for the value of *Y*. In such a case, you should replace *X* with another independent variable in the regression equation.

To determine whether $\beta_1$ is different from 0, you need to conduct a *hypothesis test*. (You find more about hypothesis testing in Chapter 12.) The name of the hypothesis test used in this case is the *t-test*, because the test statistic and critical values are based on the Student's t-distribution (covered in Chapter 11). You use this test to determine whether the slope coefficient ($\beta_1$) of the estimated regression equation is significantly different from 0. If $\beta_1 = 0$, *X* doesn't explain the value of *Y,* and the regression results are then meaningless.

The t-test is conducted in several stages. These are detailed in the following sections.

## Null and alternative hypotheses

The first is to specify the null hypothesis and the alternative hypothesis. A null hypothesis is a statement that is assumed to be true unless you find very strong evidence against it. An alternative hypothesis is a statement that is accepted instead of the null hypothesis if you reject the null hypothesis.

With the t-test, the null hypothesis is that the slope coefficient ($\beta_1$) equals 0: $H_0 : \beta_1 = 0$.

This hypothesis implies that the independent variable (X) doesn't explain the value of the dependent variable (Y). The t-test is a very conservative test; the burden of proof is to show that X *does* explain Y.

The alternative hypothesis is that the slope coefficient doesn't equal 0: $H_1 : \beta_1 \neq 0$.


As discussed in Chapter 12, this type of alternative hypothesis is known as a *two-tailed* test.

## Level of significance

The level of significance of a hypothesis test is a measure of the likelihood of a specific type of error, known as a *Type I error*. This occurs when the null hypothesis is incorrectly rejected when it is actually true. A Type II error results when the null hypothesis is *not* rejected even though it is false. With a small level of significance, there is a very low chance of committing a Type I

error, but a relatively large probability of committing a Type II error. As the level of significance is increased, the probability of a Type I error increases but the probability of a Type II error decreases.

The choice of level of significance is based on the importance of avoiding Type I errors. When you test hypotheses about regression coefficients, the level of significance ($\alpha$) is often 0.05 (5 percent).

### *Test statistic*

A test statistic is a numerical value that is used to determine if the null hypothesis should be rejected. If the test statistic has a large value (positive or negative), the likelihood that the null hypothesis is rejected is also large.

For testing hypotheses about $\beta_1$ the appropriate test statistic is

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

This expression is known as a *t-statistic* because it follows the Student's t-distribution (covered in Chapter 11).

The term $s_{\hat{\beta}_1}$ is the *standard error* of $\hat{\beta}_1$ which you can think of as the standard deviation of $\hat{\beta}_1$. (Standard errors are covered in Chapter 10.)

In other words, $s_{\hat{\beta}_1}$ is the amount of *uncertainty* associated with the use of $\hat{\beta}_1$ to estimate $\beta_1$. The larger the standard error of $\hat{\beta}_1$, the less likely you are to reject the null hypothesis that $\beta_1 = 0$.

You compute the test statistic for this hypothesis test as follows:

Also known as *standard error of the regression (SER)*, the SEE is a measure of the dispersion of the sample values above and below the estimated regression line.

$$SEE = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2}}$$

Based on Table 15-2, SEE is computed as follows:

RSS is computed as:

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = 0.14$$

With a sample size of 8, SEE equals:

$$SEE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2}}$$
$$= \sqrt{\frac{0.14}{6}}$$
$$= 0.15275$$

1. **Calculate the standard error of $\hat{\beta}_1$:**

$$S_{\hat{\beta}_1} = \frac{SEE}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}}$$

SEE equals 0.15275. $\sum_{i=1}^{n} X_i^2$ represents the sum of the squared values of X. $n\bar{X}^2$ represents the sample size times the square of the sample mean.

You can get the values of $\sum_{i=1}^{n} X_i^2$ and $n\bar{X}^2$ from Table 15-3.

**Table 15-3 The Standard Error of $\hat{\beta}_1$**

| $X_i$ | $X_i^2$ |
|---|---|
| 16 | 256 |
| 14 | 196 |
| 12 | 144 |
| 11 | 121 |
| 12 | 144 |
| 15 | 225 |
| 13 | 169 |
| 11 | 121 |

The sample mean is obtained by adding up the values in the $X_i$ column, then dividing the sum by the sample size of 8:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{16+14+12+11+12+15+13+11}{8} = 13$$

The sum of the squared values of X is obtained by squaring each value of X and then summing the results:

$$\sum_{i=1}^{n} X_i^2 = 256 + 196 + 144 + 121 + 144 + 225 + 169 + 121 = 1{,}376$$

The formula for computing the standard error of $\hat{\beta}_1$ is:

$$S_{\hat{\beta}_1} = \frac{SEE}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}} = \frac{0.15275}{\sqrt{1{,}376 - (8)(13^2)}} = 0.03118$$

2. **Calculate the test statistic:**

   $\hat{\beta}_1 = 0.15$ (see the earlier section "Estimating the Population Regression Equation"); therefore, combining this with the standard error of $\hat{\beta}_1$, the t-statistic for $\hat{\beta}_1$ is computed as

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{0.15}{0.03118} = 4.81$$

### Critical values

A critical value shows the number of standard deviations away from the mean of a distribution where:

- a specified percentage of the distribution is above the critical value
- the remainder of the distribution is below the critical value

To test a hypothesis, the test statistic is compared with one or two critical values. If the test statistic is more *extreme* than the relevant critical value, the null hypothesis is rejected. Otherwise, the null hypothesis fails to be rejected. (It's technically incorrect to say that a null hypothesis is accepted, because you don't know every value in the population being tested.)

With simple regression analysis, the critical values are taken from the Student's t-table with $n - 2$ degrees of freedom. (These are found in Table 15-4.)


Degrees of freedom refers to the number of *independent* values in a sample. When it's necessary to estimate two measures from a sample (in this case, $\hat{\beta}_0$ and $\hat{\beta}_1$) the number of degrees of freedom equals the sample size minus 2.


The Student's t-distribution is a continuous distribution that has a mean of zero and a larger variance and standard deviation than the standard normal distribution (covered in Chapter 9). The distribution is sometimes described as having "fat tails" because it's more spread out.

The purpose of the t-distribution is to describe the statistical properties of sample means that are estimated from *small* samples; the standard normal distribution is used for *large* samples. (There's much more about the Student's t-distribution Chapter 11.)

In this case, say you choose the level of significance ($\alpha$) to be 0.05. (This is a widely used value for testing hypotheses about regression coefficients.) Because the sample size ($n$) is 8, the appropriate critical values are

$$\pm t_{\alpha/2}^{n-2} = \pm t_{0.025}^{6}$$

You find these values in the Student's t-table, such as Table 15-4.

**Table 15-4 The Student's t-Distribution**

| *Degrees of Freedom* | *t0.10* | *t0.05* | *t0.025* | *t0.01* | *t0.005* |
|---|---|---|---|---|---|
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |

You find the value of the positive critical value $t_{0.025}^{6}$ at the intersection of the row for 6 degrees of freedom and the column labeled $t_{0.025}$, which is 2.447. The value of the negative critical value $-t_{0.025}^{6}$ is then –2.447.

### Decision rule

A decision rule is used to determine if the null hypothesis should be rejected. Because the alternative hypothesis is $H_1 : \beta_1 \neq 0$, there are two critical values: one positive, one negative. (These are shown to be -2.447 and 2.447 in the previous section.)

If the test statistic is either greater than 2.447 or less than -2.447, the null hypothesis will be *rejected*. This indicates that there is strong evidence that the slope coefficient $\beta_1$ is not equal to zero; in other words, the regression equation *does* explain the relationship between the dependent variable (GPA) and the independent variable (monthly hours of studying).

If the test statistic falls between these two values, the null hypothesis *fails* to be rejected. In this case, there is insufficient evidence to reject the hypothesis

that $\beta_1$ equals zero. This shows that the regression equation does *not* explain the relationship between the dependent variable (GPA) and the independent variable (monthly hours of studying).

In this case, the test statistic is 4.81, which is greater than 2.447. Therefore, you reject the null hypothesis in favor of the alternative hypothesis, indicating that $\hat{\beta}_1$ is different from 0 (that is, it's *statistically significant*). Therefore, strong evidence shows that *X* (monthly hours of studying) does explain the value of *Y* (GPA.)

This result does not imply that hours of studying is the *only* factor that explains GPA, but it is an important determinant of GPA.

You can also test whether the estimated intercept ($\hat{\beta}_0$) is statistically significant, but often doing so isn't necessary. The slope coefficient is the most important value in the regression equation.

# Using Statistical Software

Many spreadsheet programs (such as Excel) and specialized statistical packages (such as SPSS) allow you to generate the results you need for regression analysis. For example, you can use a spreadsheet program to get the results shown in Figure 15-9 for the GPA example from the "Estimating the Population Regression Equation" section earlier in this chapter (these results were generated using Excel).

As you can see, Figure 15-9 shows the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ under the *Coefficients* column; the values of the coefficient of determination ($R^2$) and the standard error of the estimate, under the *Regression Statistics* column; and the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ and the t-statistics, under the columns *Standard Error* and *t-Stat*.

Figure 15-9 provides one additional useful measure you can use to test hypotheses about the coefficients, called *p-values* (or *probability values*). The p-value represents the likelihood of finding the given t-statistic if the null hypothesis is true. An extremely low p-value indicates that the null hypothesis of a 0 coefficient should be rejected. More formally, when testing the hypothesis $H_0 : \beta_1 = 0$, if the p-value is less than the level of significance ($\alpha$), the null hypothesis is rejected; otherwise, it isn't rejected.

In this example, the p-value for $\hat{\beta}_1$ is 0.002968105; the level of significance is 0.05; therefore, because the p-value is less than the level of significance, the null hypothesis is rejected, confirming the results found when testing the hypothesis with the t-statistic.

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.891132789 |
| R Square | 0.794117647 |
| Adjusted R Square | 0.759803922 |
| Standard Error | 0.152752523 |
| Observations | 8 |

ANOVA

| | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Regression | 1 | 0.54 | 0.54 | 23.14285714 |
| Residual | 6 | 0.14 | 0.023333333 | |
| Total | 7 | 0.68 | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
|---|---|---|---|---|
| Intercept | 1.05 | 0.408928138 | 2.56768831 | 0.042466896 |
| X (Monthly Hours) | 0.15 | 0.031180478 | 4.810702354 | 0.002968105 |

**Figure 15-9:** GPA regression problem.


Using the t-statistic or the p-value to test the significance of a regression coefficient will always provide the same results.

# Assumptions of Simple Linear Regression

The simple regression model shown in this chapter is based on several extremely important assumptions. If any of these assumptions are violated, the reliability of the regression results is questionable.

The most important assumptions include the following:

- The expected value of each error term is 0; that is, $E(\varepsilon_i) = 0$. So although some error terms are positive and some are negative, they add up to 0.
- The variances of the error terms are finite and constant for all values of $x_i$; this common variance is designated $\sigma^2$.
- The error terms are independent of each other (for example, they don't influence each other).
- Each error term, $\varepsilon_i$, is independent of the corresponding value of $X_i$ (the value of X*i* doesn't influence the value of the error term and vice versa).

- The error terms are *normally distributed*. Although this assumption isn't required for linear regression, it's often used and allows you to compute confidence intervals for the regression coefficients. It also allows you to test hypotheses about the coefficients.

With simple regression analysis, two of the most important violations of the assumptions include autocorrelation and heteroscedasticity:

- **Autocorrelation** occurs when the error terms are correlated with each other (they are related to each other). It violates the assumption of independence. Two independent variables have a correlation of 0 between them.

  Autocorrelated error terms can cause the standard errors of the regression coefficients to be understated, thus increasing the risk that coefficients will be incorrectly found to be statistically significant (for example, different from zero).

- **Heteroscedasticity** occurs when the error terms don't have a constant variance. This problem can cause the standard errors of the regression coefficients to be understated, increasing the risk that coefficients will be incorrectly found to be statistically significant (for example, different from zero).


Formal statistical tests are available to help you determine whether these problems are present. For example, the Durbin-Watson test is used to find evidence of autocorrelation in sample data. (More details can be found at `http://en.wikipedia.org/wiki/Durbin%E2%80%93Watson_statistic`.) The White test is used to find evidence of heteroscedasticity in sample data. (More details can be found at `http://en.wikipedia.org/wiki/White_test`.)

If autocorrelation is present, you may use the Cochrane-Orcutt procedure, which adjusts the regression model for autocorrelation. (More details can be found at `https://en.wikipedia.org/wiki/Cochrane-Orcutt_procedure`.)

In the case of heteroscedasticity, you may transform the variables into natural logarithms and rerun the regression equation; for example, the dependent variable could be ln(Y) and the independent variable could be ln(X). ("ln" is standard for natural logarithm.) More formal procedures are also available to correct for heteroscedasticity, such as heteroscedasticity-consistent standard errors. (More information about this procedure is found at `http://en.wikipedia.org/wiki/Heteroscedasticity-consistent_standard_errors`.)

# Chapter 16

# Multiple Regression Analysis: Two or More Independent Variables

### In This Chapter

- Getting familiar with the assumptions underlying multiple regression analysis
- Implementing the multiple regression model
- Watching for multicollinearity

You use regression analysis to estimate the strength and direction of the relationship between two or more variables. As I explain in Chapter 15, simple regression analysis allows you to estimate the relationship between a dependent variable ($Y$) and an independent variable ($X$).

In this chapter, I explore the possibilities of multiple regression analysis, which you use to estimate the relationship between a dependent variable ($Y$) and two or more independent variables ($X_1, X_2, \ldots$).

The additional independent variable(s) introduces more complications into multiple regression analysis. In particular, it takes more statistical testing to validate the results of a multiple regression model. Further, additional errors may arise in multiple regression analysis that can't take place with simple regression analysis.

This chapter explains how to implement multiple regression analysis, how to test the results, and what potential pitfalls may arise.

# The Fundamental Assumption: Variables Have a Linear Relationship

Just as with simple regression analysis (discussed in Chapter 15), multiple regression analysis is based on the assumption that the dependent variable and the independent variables have a *linear relationship* between them.

If two variables are linearly related, a graph of their relationship is a straight line. The equation of a straight line is:

$$Y = mX + b$$

- $X$ is the independent variable
- $Y$ is the dependent variable
- $m$ is the slope coefficient
- $b$ is the intercept

In this equation, the value of Y depends on the value of X (not the other way around). The slope tells you *how much* Y changes when X changes; the intercept tells you the value of Y when X equals 0.

For example, suppose that the equation of a straight line is:

$Y = 4X - 7$

The slope of 4 shows that:

- if X increases by 1, Y increases by 4
- if X decreases by 1, Y decreases by 4

The intercept of –7 shows that Y equals –7 when X equals 0.

In addition to having a linear relationship between the dependent variable and each independent variable, there must be a joint linear relationship between the dependent variable and *all* the dependent variables.

If variables don't have a linear relationship, you can still use regression analysis; however, you may have to make adjustments to the regression equation. For example, it may be that the relationship between $Y$ and $X$ is *nonlinear* but that the relationship between $\ln(Y)$ — the *natural logarithm* with base $e = 2.71828$ — and $X_1$ and $X_2$ is linear. In this case, you can run a regression using $\ln(Y)$ as the dependent variable and $X_1$ and $X_2$ as the independent variables. Another possibility is that the relationship between $\ln(Y)$, $\ln(X_1)$, and $\ln(X_2)$ is linear. In this case, you use $\ln(Y)$ as the dependent

variable, and $\ln(X_1)$ and $\ln(X_2)$ as the independent variables. (Logarithmic transformations for regression analysis are discussed in Chapter 15.)

# Estimating a Multiple Regression Equation

With multiple regression analysis, the population regression equation may contain any number of independent variables, such as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

In this case, there are $k$ independent variables, indexed from 1 to $k$.

For example, suppose that the Human Resources department of a major corporation wants to determine whether the salaries of its employees are related to the employees' years of work experience and their level of graduate education. To test this idea, the HR department picks a sample of eight employees randomly and records their annual salaries (measured in thousands of dollars per year), years of experience, and years of graduate education.

The following variables are defined:

- $Y$ represents an employee's annual salary, measured in thousands of dollars.
- $X_1$ represents an employee's number of years of job experience. A value of 0 represents someone who has no job experience (such as a recent college graduate).
- $X_2$ represents the number of years of graduate education. A value of 0 represents a college graduate with no graduate education.

The following lists the sample data.

| *Y (Annual Salary, in Thousands)* | *$X_1$ (Years of Experience)* | *$X_2$ (Years of Graduate Education)* |
|---|---|---|
| 80 | 1 | 0 |
| 90 | 2 | 1 |
| 100 | 3 | 2 |
| 120 | 4 | 2 |
| 85 | 1 | 0 |
| 95 | 2 | 1 |
| 105 | 2 | 2 |
| 140 | 8 | 3 |

The HR department runs a regression using a spreadsheet program, such as Excel. Figure 16-1 shows the results.

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.971774936 |
| R Square | 0.944346527 |
| Adjusted R Square | 0.922085137 |
| Standard Error | 5.52943278 |
| Observations | 8 |

ANOVA

| | *df* | *SS* | *MS* | *F-stat* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 2 | 2594.001866 | 1297.000933 | 42.42082621 | 0.000730686 |
| Residual | 5 | 152.8731343 | 30.57462687 | | |
| Total | 7 | 2746.875 | | | |

| | *Coefficients* | *Standard Error* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|
| Intercept | 76.47014925 | 3.397844048 | 22.50549118 | 3.21898E-06 | 85.20458544 |
| $X_1$ (Years of experience) | 5.320895522 | 1.695561146 | 3.13813249 | 0.025720437 | 9.679474206 |
| $X_2$ (Years of graduate education) | 7.350746269 | 3.669054725 | 2.003444162 | 0.101492144 | 16.7823517 |

**Figure 16-1:** Spreadsheet showing salary regression results.

Taking the intercept and slope coefficients ($X_1$ and $X_2$) from the *Coefficients* column in Figure 16-1, you can fill in the estimated regression equation as

$$\hat{Y}_i = 76.47 + 5.32X_{1i} + 7.35X_{2i}$$

(The values are rounded to two decimal places.)

This equation shows that the following is true for this firm:

- The starting salary for a new employee with no experience or graduate education is $76,470. This amount is based on the intercept of the regression equation.
- Each additional year of experience adds $5,320 to an employee's salary; this amount is based on the coefficient of $X_1$ (years of experience).
- Each additional year of graduate education adds $7,350 to an employee's salary, which is based on the coefficient of $X_2$ (years of graduate education).


In each case, you multiply the coefficients by $1,000 to get the impact on salary because these variables are measured in thousands of dollars per year.

The intercept of the equation, 76.47, shows the value of $Y$ (the employee's annual salary) when *both* $X_1$ (years of experience) and $X_2$ (years of graduate education) equal 0 (that is, a new employee with no experience or graduate education). The intercept shows that the starting salary is $76.47 \times \$1{,}000 = \$76{,}470$.

The coefficient of $X_1$, 5.32, shows how much $Y$ changes due to a one-unit change in $X_1$. Because $X_1$ represents years of experience, a one-unit change in $X_1$ is one additional year of experience. Therefore, each additional year of experience adds $5.32 \times \$1{,}000 = \$5{,}320$ to an employee's salary.

The coefficient of $X_2$, 7.35, shows how much $Y$ changes due to a one-unit change in $X_2$. Because $X_2$ represents years of graduate education, a one-unit change in $X_2$ is one additional year of graduate school. Therefore, each additional year of graduate school adds $7.35 \times \$1{,}000 = \$7{,}350$ to an employee's salary.

The following sections show how you can use the results from the spreadsheet to predict the salary of an employee with a specified number of years of experience and education. A new measure is introduced to determine how well the regression equation "fits" or explains the sample data; this is known as the *adjusted coefficient of determination*.

Two types of hypothesis tests are covered. A hypothesis is tested for all the slope coefficients of the regression equation as a group; if this hypothesis fails to be rejected, then *none* of the independent variables belong in the regression equation. Hypotheses are also tested about the individual slope coefficients of the regression equation to see if any of the independent variables should be discarded from the regression equation.

# Predicting the value of Y

You can use the multiple regression equation for employee salaries from the previous section to predict the annual salary of an employee with a specific amount of experience and education. For example, suppose that a randomly chosen employee has five years of experience and one year of graduate education. The predicted salary of this employee is

$$\hat{Y}_i = 76.47 + 5.32X_{1i} + 7.35X_{2i}$$
$$\hat{Y}_i = 76.47 + 5.32(5) + 7.35(1)$$
$$\hat{Y} = 110.42$$

This result shows that the predicted annual salary is (110.42)($1,000) = $110,420.

## The adjusted coefficient of determination

You can use several methods to test how well a multiple regression equation actually fits, or explains, the relationship between a dependent variable and one or more independent variables in a given data set. One of these methods is to use the *adjusted coefficient of determination* to determine how well the regression equation "fits" the sample data. The adjusted coefficient of determination is closely related to the coefficient of determination (also known as $R^2$) you use to test the results of a simple regression equation (shown in Chapter 15).

The adjusted coefficient of determination (also known as adjusted $R^2$ or ($\bar{R}^2$, pronounced "R bar squared") is a statistical measure that shows the proportion of *variation* explained by the estimated regression line.

*Variation* refers to the sum of the squared differences between the values of $Y$ and the mean value of $Y$, expressed mathematically as

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$$

Adjusted $R^2$ always takes on a value between 0 and 1. The closer adjusted $R^2$ is to 1, the better the estimated regression equation fits or explains the relationship between $X$ and $Y$.

The key difference between $R^2$ and adjusted $R^2$ is that $R^2$ increases automatically as you add new independent variables to a regression equation (even if they don't contribute any new explanatory power to the equation). Therefore, you want to use adjusted $R^2$ with multiple regression analysis. Adjusted $R^2$ increases only when you add new independent variables that *do* increase the explanatory power of the regression equation, making it a much more useful measure of how well a multiple regression equation fits the sample data than $R^2$.

The following equation shows the relationship between adjusted $R^2$ and $R^2$:

$$\bar{R}^2=1-\left(1-R^2\right)\left[\frac{n-1}{n-(k+1)}\right]$$

- n = the sample size
- k = the number of independent variables in the regression equation

Figure 16-2 highlights a section of the regression statistics from the spreadsheet in Figure 16-1.

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.971774936 |
| R Square | 0.944346527 |
| Adjusted R Square | 0.922085137 |
| Standard Error | 5.52943278 |
| Observations | 8 |

**Figure 16-2:** Spreadsheet showing the adjusted coefficient of determination.

Figure 16-2 shows the adjusted coefficient of determination (*Adjusted R Square*) as approximately 0.922.This is computed as follows:

$$\bar{R}^2 = 1-\left(1-R^2\right)\left[\frac{n-1}{n-(k+1)}\right]$$

$R^2$ is found on Figure 16-2; it's labeled "R Square" and equals 0.944346527. Because the sample contains eight observations, and there are two independent variables (years of experience and years of graduate education), the adjusted $R^2$ is computed as:

$$\bar{R}^2 = 1-\left(1-R^2\right)\left[\frac{n-1}{n-(k+1)}\right]$$

$$= 1-\left(1-0.944346527\right)\left[\frac{8-1}{8-(2+1)}\right]$$

$$= 0.922085138$$

(This equals the value in Figure 16-2 except for a slight rounding difference.)

The range of possible values for the adjusted coefficient of determination is from 0 to 1; in mathematical terms,

$$0 \le \bar{R}^2 \le 1$$

Based on the value of adjusted $R^2$, the proportion of *variation* explained by the estimated regression line is approximately 0.922 or 92.2 percent.

## The F-test: Testing the joint significance of the independent variables

The F-test is a special type of hypothesis test that is used to determine if the independent variables in a multiple regression equation jointly determine the value of the dependent variable. This is done by testing the hypothesis that *all slope coefficients equal 0*. If true, the regression equation doesn't explain the relationship between the dependent and the independent variables. In this case, you may use a new set of independent variables to try to explain the value of the dependent variable.

In the following sections, the steps required to carry out the F-test are explained in detail, based on the salaries example found in the section "Estimating a Multiple Regression Equation." This procedure begins with the statement of the null and alternative hypotheses, along with the choice of a level of significance. The next step is to construct the test statistic and compare it to a critical value before making a decision as to the validity of the regression equation. (Hypothesis testing is introduced in Chapter 12.)

### The null and alternative hypotheses for the F-test

The first step in a hypothesis test is to specify the null hypothesis and the alternative hypothesis. A null hypothesis is a statement that is assumed to be true unless you find very strong evidence against it. An alternative hypothesis is a statement that is accepted instead of the null hypothesis if you reject the null hypothesis.

For the F-test with two independent variables, the null hypothesis is

$$H_0 : \beta_1 = \beta_2 = 0$$

This null hypothesis indicates that both slope coefficients ($X_1$ and $X_2$) equal 0. A coefficient of 0 suggests that an independent variable doesn't explain the value of the dependent variable. If you can't reject this hypothesis, then you can't use the regression equation to explain the relationship between the dependent variable (salaries) and the independent variables (years of experience and graduate education).

The alternative hypothesis is that at least one slope coefficient doesn't equal 0. In other words, at least one of the independent variables does belong in the regression equation because it explains the value of the dependent variable.

### The level of significance for the F-test

The level of significance specifies the probability of a *Type I error*. This occurs when the null hypothesis is incorrectly rejected when it is actually true. A Type II error results when the null hypothesis is *not* rejected even though it is false. In many business applications, the level of significance is chosen to be 0.01, 0.05, or 0.10, and 0.05 is a common choice.

The Greek letter $\alpha$ ("alpha") is normally used to represent the level of significance. The choice of the level of significance depends on how important it is to avoid a Type I error compared with the importance of avoiding a Type II error. The higher the level of significance, the greater is the probability of a Type I error, and the lower is the probability of a Type II error.


It's impossible to reduce the probability of *both* a Type I and a Type II error without increasing the size of the sample used to test the null hypothesis.

### The test statistic for the F-test

A test statistic is a numerical value that's used to determine if the null hypothesis should be rejected. If the test statistic has a large value (positive or negative), the likelihood that the null hypothesis will be rejected is also large.

You compute the test statistic (also known as the *F-statistic*) with this equation:

$$F = \frac{R^2 / k}{\left(1 - R^2\right) / \left[n - (k+1)\right]}$$

This test statistic is known as the F-statistic because probabilities for this statistic may be computed from the F-distribution. (The F-distribution is introduced in Chapter 13.)

In the salaries example in section "Estimating a Multiple Regression Equation," the F-statistic equals

$$\begin{aligned} F &= \frac{R^2 / k}{\left(1 - R^2\right) / \left[n - (k+1)\right]} \\ &= \frac{0.944346527 / 2}{\left(1 - 0.944346527\right) / \left[8 - (2+1)\right]} \\ &= 42.42 \end{aligned}$$

The value of $R^2$ is taken from Figure 16-2 (it is labeled "R Square"). n equals 8 because there are eight elements in the sample. k equals 2 because there are two independent variables (years of experience and years of graduate education).


The test statistic follows the *F-distribution* with $k$ numerator degrees of freedom and $[n - (k + 1)]$ denominator degrees of freedom. The F-distribution is characterized by two different types of degrees of freedom; these are known as *numerator* degrees of freedom and *denominator* degrees of freedom.

For the F-test, you can find probabilities for the test statistic from an F-table based on the level of significance, the number of numerator degrees of freedom, and the number of denominator degrees of freedom. (See Chapters 13 and 14 for more on the F-distribution and the F-table.)

Figure 16-3 shows a portion of Figure 16-1, highlighting the ANOVA (analysis of variance) table. Here, you see that the value of the F-statistic is 42.42082621, which is approximately equal to 42.42 (found under the *F*-stat column). Note that you can also obtain the value of the F-statistic from two values in the ANOVA table:

1. MS(Regression), which equals 1297.000933 and is found at the intersection of the row labeled "Regression" and the column labeled "MS"
2. MS(Residual), which equals 30.57462687 and is found at the intersection of the row labeled "Residual" and the column labeled "MS"

The ratio of these two values = 1297.000933 / 30.57462687 = 42.42082621, or approximately 42.42.

**Figure 16-3:** The ANOVA table for the salaries example.

ANOVA

| | *df* | *SS* | *MS* | *F-stat* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 2 | 2594.001866 | 1297.000933 | 42.42082621 | 0.000730686 |
| Residual | 5 | 152.8731343 | 30.57462687 | | |
| Total | 7 | 2746.875 | | | |

$R^2$ is the ratio of SS(Regression) to SS(Total). Adjusted $R^2$ is obtained from $R^2$ as

$$\bar{R}^2 = 1 - \left(1 - R^2\right)\left[\frac{n-1}{n-(k+1)}\right]$$

where n = the sample size, and k = the number of independent variables in the regression equation.

### The critical value for the F-test

A critical value shows the number of standard deviations away from the mean of a distribution where a specified percentage of the distribution is above the critical value, and the remainder of the distribution is below the critical value.

In general, when testing a hypothesis the test statistic is compared with one or two critical values. If the test statistic is more *extreme* than the relevant critical value, the null hypothesis is rejected. Otherwise, the null hypothesis fails to be rejected.

For the F-test, there's a single critical value, which is uniquely determined by the level of significance and the numerator and denominator degrees of freedom.

For the F-test, the numerator and denominator degrees of freedom are computed as follows:

- Numerator degrees of freedom: $k = 2$
- Denominator degrees of freedom: $[n - (k + 1)] = (8 - [2 + 1]) = 5$

You can choose the appropriate critical value from an F-table. (The F-table is introduced in Chapter 13; the values in the table are taken from the F-distribution.)

Unlike the tables used for most other probability distributions, you need one entire F-table for each level of significance. Table 16-1 shows an excerpt from the F-table for a 5 percent level of significance ($\alpha = 0.05$):

**Table 16-1 A Section of the F-Table with $\alpha = 0.05$**

| $\upsilon_2 \backslash \upsilon_1$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| 2 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.38 |
| 3 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 |
| 4 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 |
| 5 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 |
| 6 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 |
| 7 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 |
| 8 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 |
| 9 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 |

The top row represents the numerator degrees of freedom ($\upsilon_1$); the first column represents the denominator degrees of freedom ($\upsilon_2$).

In this example, you're looking for a right-tail area of 5 percent with $\upsilon_1 = 2$, and $\upsilon_2 = 5$. You find this critical value at the intersection of the column labeled *2* and the row labeled *5.* You express this critical value mathematically as

$$F_{\alpha}^{\upsilon 1,\upsilon 2} = F_{0.05}^{2,5} = 5.79$$

### The decision for the F-test

If the test statistic exceeds the critical value, you reject the null hypothesis; otherwise, you don't reject it. In this case, the test statistic is approximately 42.42, and the critical value is 5.79. Therefore, you reject the hypothesis that all the slope coefficients ($\beta_1$ and $\beta_2$) are equal to zero. In other words, one (or both) of the independent variables (years of experience and years of graduate education) explains the annual salaries of the employees at this company.

### Testing the null hypothesis with the p-value when testing the joint significance of the slope coefficients

As an alternative to comparing the F-statistic with a critical value, you can test the hypothesis by comparing the *p-value* (probability value) with the level of significance.

The p-value represents the probability that a test statistic will equal a specified value when the null hypothesis is true. A low p-value is evidence against a null hypothesis.

When you're using the p-value, the decision rule is this: If the p-value is less than the level of significance, you reject the null hypothesis; otherwise, you won't reject the null hypothesis.

In this example, the level of significance is 0.05 (5 percent). Figure 16-3 shows the p-value (under the *Significance F* column) as (approximately) 0.0007. Because the p-value is well below the level of significance, you reject the null hypothesis. Therefore, at least one of the slope coefficients is statistically significant at the 5 percent level.

## The t-test: Determining the significance of the slope coefficients

After you use the F-test to confirm that at least one slope coefficient isn't equal to 0, you test each slope coefficient separately to determine if it belongs in the regression equation; this requires the use of a hypothesis test

known as the t-test. (The test has this name because the test statistic and the critical values are taken from the Student's t-distribution. See Chapter 15 for more on this test.) The t-test lets you determine which of the slope coefficients is statistically significant or if both are statistically significant.

### Null and alternative hypotheses for the t-test

With the t-test, the null hypothesis states that a slope coefficient equals 0. For example, to test the hypothesis that $\beta_1 = 0$, you would write the null hypothesis as $H_0 : \beta_1 = 0$.

There are three possible alternative hypotheses:

- $H_1 : \beta_1 > 0$: This is known as a *right-tailed* test
- $H_1 : \beta_1 < 0$: This is known as a *left-tailed* test
- $H_1 : \beta_1 \neq 0$: This is known as a *two-tailed* test

With a right-tailed test, you are looking for evidence that the actual value of $\beta_1$ is *greater than* 0; with a left-tailed test, you are looking for evidence that the actual value of of $\beta_1$ is *less than* 0. With a two-tailed test, you are looking for evidence that the actual value of $\beta_1$ is *different from* 0. For the t-test, the two-tailed approach is usually taken.

### Level of significance for the t-test

When you test hypotheses about individual regression coefficients, the level of significance ($\alpha$) is often set equal to 0.05 (5 percent). Other commonly used choices include 0.001, 0.01, 0.05, and 0.10.

### Test statistic for the t-test

For the t-test, the test statistic is the ratio of the estimated coefficient to the standard error of the coefficient. For example, the test statistic for determining whether $\beta_1 = 0$ is

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$


This expression is known as a *t-statistic* because it follows the t-distribution. (You can compute probabilities for the t-statistic from a Student's t-table. See Chapter 11 for more discussion of the Student's t-distribution.)

You can find the values you need to construct the t-statistic from the regression statistics under the *Coefficients* and *Standard Error* columns, as shown in Figure 16-4.

**Figure 16-4:** Coefficients and standard errors for the salary example.

| | *Coefficients* | *Standard Error* | *t-Stat* | *P-value* |
|---|---|---|---|---|
| Intercept | 76.47014925 | 3.397844048 | 22.50549118 | 3.21898E-06 |
| $X_1$ (Years of experience) | 5.320895522 | 1.695561146 | 3.13813249 | 0.025720437 |
| $X_2$ (Years of graduate education) | 7.350746269 | 3.669054725 | 2.003444162 | 0.101492144 |

As you can see in Figure 16-4, for the variable $X_1$ (years of experience), the coefficient is (approximately) 5.32, and the standard error is (approximately) 1.70.

The ratio of these two values is

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{5.32}{1.70} = 3.13$$

Figure 16-4 also shows that for the variable $X_2$ (years of graduate education), or that $\beta_2 = 0$, the coefficient is (approximately) 7.35, and the standard error is (approximately) 3.67.

The ratio of these two values is

$$t = \frac{\hat{\beta}_2}{s_{\hat{\beta}_2}} = \frac{7.35}{3.67} = 2.00$$

### Critical values for the t-test

With a multiple regression equation, you take the critical values from the Student's t-table with $n - (k + 1)$ degrees of freedom ($n$ is the sample size and $k$ is the number of independent variables).

The number of degrees of freedom refers to the number of *independent* elements in a sample.


When testing hypotheses about a slope coefficient, the degrees of freedom equals the sample size (n) minus k+1 (k is the number of independent variables in the regression equation). This is because the sample data is used to estimate k+1 values: These are the estimated intercept and k estimated slope coefficients. As a result, the degrees of freedom equal n–(k+1).

The critical value depends on the alternative hypothesis as follows:

- For a right-tailed test, there is a single critical value, $+t_{\alpha}^{n-(k+1)}$

  If the test statistic is *greater than* this value, the null hypothesis is *rejected*; otherwise, it fails to be rejected.

- For a left-tailed test, there is a single critical value, $-t_{\alpha}^{n-(k+1)}$

  If the test statistic is *less than* this value, the null hypothesis is *rejected*; otherwise, it fails to be rejected.

- For a two-tailed test, there are two critical values, $\pm t_{\alpha/2}^{n-(k+1)}$

  If the test statistic is *greater than* the positive critical value or *less than* the negative critical value, the null hypothesis is *rejected*; otherwise, it fails to be rejected.

When testing hypotheses about the slope coefficients in a regression equation, the appropriate number of degrees of freedom equals n – (k + 1); n is the sample size and k is the number of independent variables. For the salaries example, the sample size is 8 and there are two independent variables (years of experience and years of graduate education.) Therefore, the degrees of freedom equals $n - (k + 1) = 8 - (2 + 1) = 5$.

Because this is a two-tailed test, two critical values occur:

$$\pm t_{\alpha/2}^{n-(k+1)} = \pm t_{0.025}^{5}$$

You can find these critical values in a Student's t-table, which is based on the Student's t-distribution (see Chapter 11 for details). Table 16-2 shows an excerpt:

**Table 16-2 The Student's t-Distribution**

| *Degrees of Freedom* | $t_{0.10}$ | $t_{0.05}$ | $t_{0.025}$ | $t_{0.01}$ | $t_{0.005}$ |
|---|---|---|---|---|---|
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |

The t-distribution (also known as the Student's t-distribution) is a continuous probability distribution that has a mean of zero, is symmetrical about its

mean, and has more areas in the "tails" of the distribution than the standard normal distribution. (The standard normal distribution is found in Chapter 9; the Student's t-distribution is found in Chapter 11.) The Student's t-distribution is uniquely characterized by its degrees of freedom.

You find the value of the positive critical value, $t^5_{0.025}$, at the intersection of the row labeled *5* degrees of freedom and the column labeled $t_{0.025}$. The positive critical value is 2.571. Due to the symmetry of the Student's t-distribution, the negative critical value equals the positive critical value with a negative sign: –2.571.

### Decision rule for the t-test

For testing the hypothesis $H_0 : \beta_1 = 0$, you reach the appropriate decision as follows:

- If the value of the test statistic is greater than 2.571, you reject the null hypothesis $H_0 : \beta_1 = 0$ in favor of the alternative hypothesis $H_1 : \beta_1 > 0$.
- If the value of the test statistic is less than –2.571, you reject the null hypothesis $H_0 : \beta_1 = 0$ in favor of the alternative hypothesis $H_1 : \beta_1 < 0$.
- If the test statistic falls between –2.571 and 2.571, you don't reject the null hypothesis $H_0 : \beta_1 = 0$.

You follow the same process when you test the hypothesis $H_0 : \beta_2 = 0$.

For $\beta_1$, the test statistic is 3.13, which is greater than 2.571. Therefore, you reject the null hypothesis $H_0 : \beta_1 = 0$ in favor of $H_1 : \beta_1 > 0$, which indicates that $\beta_1$ is different from 0 (that is, it's statistically significant). Therefore, in the example used throughout this chapter, strong evidence shows that $X_1$ (years of experience) explains some of the value of $Y$ (annual salary).

For $\beta_2$, the test statistic is 2.00, which is between –2.571 and 2.571. Therefore, you don't reject the null hypothesis $H_0 : \beta_2 = 0$. You have insufficient evidence to show that that $X_2$ (years of graduate education) explains the value of $Y$ (annual salary).

### Testing the null hypothesis with the p-value when testing the individual slope coefficients

As an alternative to comparing the t-statistic with critical values, you can test the hypothesis by comparing the p-value with the level of significance. The decision rule is then if the p-value is less than the level of significance, you reject the null hypothesis; otherwise, you don't reject the null hypothesis.

In the example of the employee salaries, the level of significance is 0.05 (5 percent). You can find the p-values for $X_1$ and $X_2$ by referring to Figure 16-4.

For $\beta_1$, the p-value is 0.025720432, which is less than 0.05. Therefore, you reject the null hypothesis $H_0 : \beta_1 = 0$ in favor of $H_1 : \beta_1 > 0$, which indicates that $\beta_1$ is different from 0 (that is, it's statistically significant).

For $\beta_2$, the p-value is 0.101492144, which is greater than 0.05. Therefore, you don't reject the null hypothesis $H_0 : \beta_2 = 0$. So $\beta_2$ is *not* different from 0 (that is, it's *not* statistically significant).


The results you get from using the p-value always match the results of comparing a test statistic with critical values.

# Checking for Multicollinearity

One of the potential difficulties with multiple regression analysis is *multicollinearity*. Multicollinearity occurs when two or more of the independent variables are highly correlated with each other, causing the correlated variables to have large standard errors, so they appear to be statistically insignificant even if they're not. (In other words, there's a risk that independent variables are removed from the regression equation that should be included.)

Multicollinearity is unique to multiple regression because it has multiple independent variables (simple regression has only one independent variable so that multicollinearity can't occur).

A statistic known as the variance inflation factor (VIF) may be used to check for multicollinearity. As a rule of thumb, if the VIF is 10 or more, this is a sign that multicollinearity is present.

One approach to removing multicollinearity is to eliminate one of the correlated variables from the regression. Doing so lowers the p-values of the uncorrelated independent variables, which reduces the risk that they'll be considered statistically insignificant when they're not.

# Chapter 17

# Forecasting Techniques: Looking into the Future

### *In This Chapter*

- Developing time series models with regression analysis
- Modeling seasonality in a time series
- Using smoothing techniques
- Determining which model produces the best forecasts

Suppose you could forecast the price of Apple stock at the end of closing tomorrow. How rich could you be? What if you could foresee the future path of interest rates? How much of an advantage would you have over other investors? Trying to predict the future is an ancient art; some would suggest that the newest mathematical techniques are no more successful than tarot cards and Ouija boards.

Despite of the difficulty of forecasting the future, economists, investors, analysts, and traders do attempt to predict future values of economic variables, such as stock prices, commodity prices, interest rates, exchange rates, and so on. Many trading strategies depend on being able to use past history to correctly forecast the future. When these strategies succeed, it's an open question whether their success was due to sophisticated models or just plain dumb luck.

While forecasting is notoriously difficult, there are several classical techniques that may be useful for short-term business forecasting. These models have one thing in common — all base their predictions on past history and the assumption that history will be repeated in the future. This chapter focuses on these techniques, which include linear trend models, quadratic trend models, seasonally adjusted models, and exponential smoothing models.

# Defining a Time Series

A *time series* is a sequence of random variables indexed by time. (Random variables are introduced in Chapter 7.) You express a time series as $\{y_t\}$, where $y_t$ is the value of a random variable at time $t$. For example, daily closing price of IBM is a random variable because its value isn't known prior to the end of the trading day. The daily closing price of IBM stock over ten trading days can be represented as $\{y_t\} = y_1, y_2, y_3, ..., y_{10}$. $y_t$ is the price of IBM stock at time $t$.

A time series may contain the following effects:

- **Trend effects** refers to a long-run increase or decrease in a time series. For example, gold prices taken from the past 40 years would show a very strong positive trend because prices have risen consistently over this period Trends may be due to a large number of different factors, such as population growth, technological improvement, and increasing incomes.
- **Seasonal effects** refer to the impact of the time of year on economic variables. For example, sales of bathing suits, surfboards, and so forth are much stronger during the warmer months. Sales of Christmas trees, turkeys, and pumpkin pies are stronger during the colder months. Many variables aren't affected by the season; for example, the price of office furniture is not likely to fluctuate due to changes in the season.
- **Cyclical effects** refer to the impact of the business cycle. For example, sales of expensive items, such as new homes and new cars, decline when the economy falls into recession. As another example, interest rates tend to fall during recessions and rebound during recoveries.
- **Irregular effects** refer to the impact of random events such as strikes, earthquakes, sudden changes in the weather, and so on.

# Modeling a Time Series with Regression Analysis

A time series may be modeled in several different ways; one of these is to use regression analysis. (Simple regression analysis is covered in Chapter 15, and multiple regression analysis is covered in Chapter 16.) In this case, the value

of the time series being modeled is assumed to depend only on the passage of time; for example, time is the independent variable.

The basic form of a time series regression model can be expressed as $y_t = TR_t + \varepsilon_t$.

$TR_t$ is the trend of the time series at time *t,* and $\varepsilon_t$ is an error term at time *t.*

To estimate a time series with regression analysis, the first step is to identify the type of trend (if any) that's present in the data. The type of trend determines the exact equation that is estimated. After this has been specified, the next step is to run a regression of the time series data using time as the independent variable. The final step is to test the validity of the results.

This section explains the different types of trends that may be encountered in time series data, such as linear trends and quadratic trends.

# Classifying trends

In the following sections, I define the basic types of trends that may appear in a time series.

## No trend

In the case where a time series doesn't increase or decrease over time, it may instead randomly fluctuate around a constant value. In this case, the time series has *no trend*. The trend equation is set equal to a constant, which is the intercept of a regression equation:

$$TR_t = \beta_0$$

The corresponding regression equation is $y_t = \beta_0 + \varepsilon_t$.

When no trend occurs, the values of the time series may rise or fall, but on average they tend to return to the same level ($\beta_0$; (for example, the intercept of the regression equation). Figure 17-1 shows a time series with no trend.

Notice that the values of *Y* are randomly rising and falling; there is no clear pattern in the data.

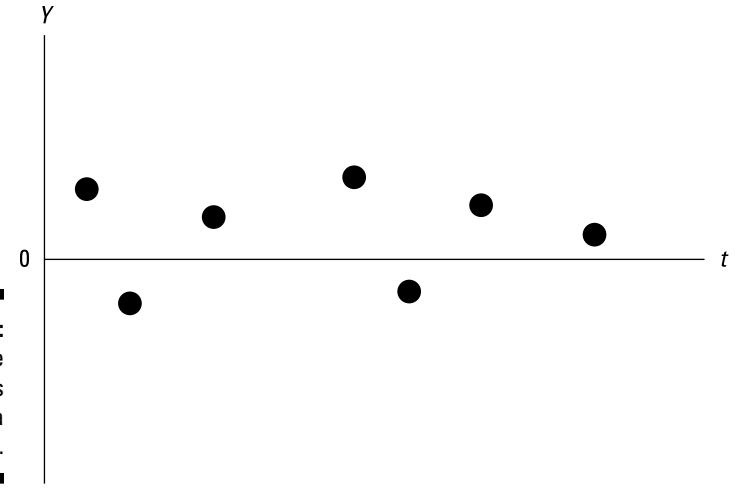



**Figure 17-1:** A time series without a trend.

### *Linear trend*

With a linear trend, the values of a time series tend to rise or fall at a constant rate ($\beta_1$). The linear trend is expressed as $TR_t = \beta_0 + \beta_1 t$.

The corresponding regression equation is $y_t = \beta_0 + \beta_1 t + \varepsilon_t$.

Figure 17-2 shows a time series with a positive linear trend. With this type of trend, the independent variable $y_t$ *increases* at a constant rate over time. (If a time series has a negative linear trend, the independent variable $y_t$ *decreases* at a constant rate over time.)

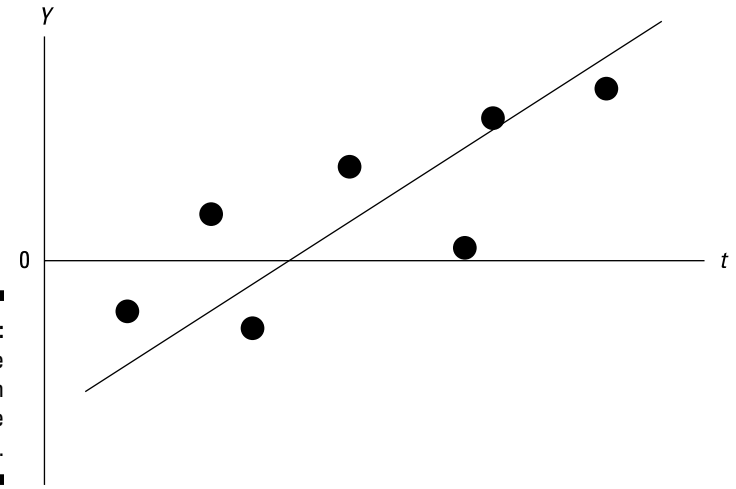



**Figure 17-2:** A time series with a positive linear trend.

Note that as *t* increases (such as time elapses), *Y* tends to increase on average. The trend line drawn through the values of *Y* has a positive slope, indicating that *Y* has a positive linear trend.

### Quadratic trend

With a quadratic trend, the values of a time series tend to rise or fall at a rate that that is not constant; it changes over time. As a result, the trend is not a straight line. The trend is expressed as $TR_t = \beta_0 + \beta_1 t + \beta_2 t^2$.

The corresponding regression equation is $y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$.

Figure 17-3 shows a time series with a quadratic trend. In this case, the value of $y_t$ increases at an increasing rate over time.

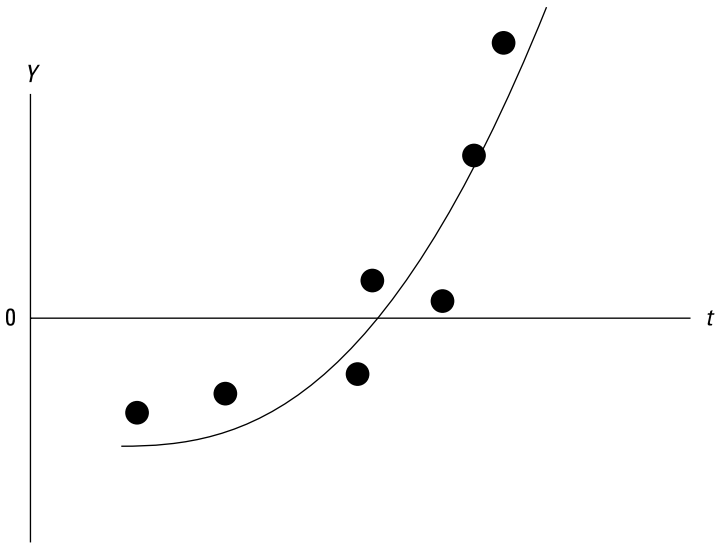



**Figure 17-3:** A time series with a quadratic trend.

Note that as *t* increases (such as time elapses), *Y* tends to increase at an increasing rate. The trend is curving upward; this type of curve indicates that the *Y* has a positive *quadratic* trend.


A quadratic equation has at least one squared term. For example, the following is a quadratic equation:

$$Y = X^2 + X + 3$$

### *Other possible trends*

It's possible that a trend may contain terms that are raised to the third power, fourth power, or higher. This type of trend is extremely rare in business applications. Most time series of financial data have a linear trend, a quadratic trend, or no trend at all.

## *Estimating the trend*

To estimate a time series, a trend must be estimated. You begin by creating a line chart of the time series (line charts are introduced in Chapter 2). The line chart shows how a variable changes over time; it can be used to inspect the characteristics of the data, in particular, to see whether a trend. For example, suppose you're a portfolio manager and you have reason to believe a linear trend occurs in a time series of returns to Microsoft stock. You plot the monthly prices from August 2008 to July 2013 on a graph like Figure 17-4.

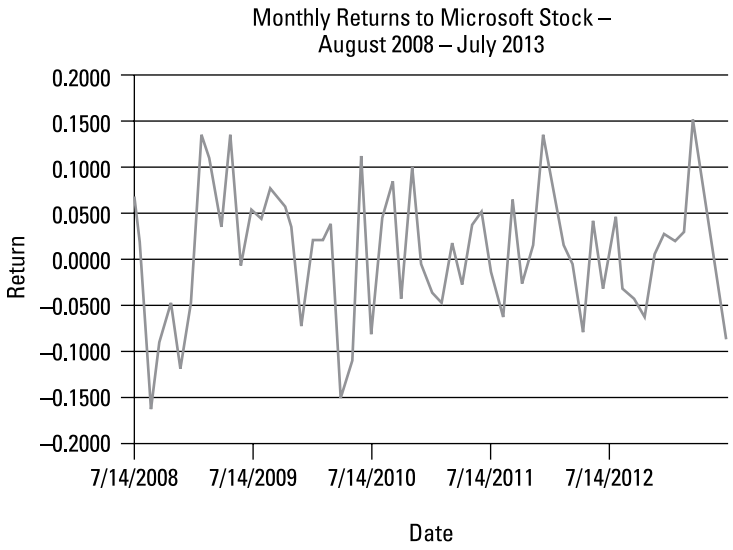



**Figure 17-4:** Monthly returns to Microsoft stock.

According to Figure 17-4, no trend occurs in the data. The returns rise and fall with no particular pattern.

To formally test whether a linear trend occurs, run a time series regression with a time trend as the independent variable, which you can set up like so:

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

In this example, the dependent variable is the price of Microsoft stock, and the independent variable is time (measured in months).

Figure 17-5 shows the results of this regression analysis.

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.052602349 |
| R Square | 0.002767007 |
| Adjusted R Square | –0.014426665 |
| Standard Error | 0.073283809 |
| Observations | 60 |

ANOVA

| | *df* | *SS* | *MS* | *F-stat* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 0.000864286 | 0.000864286 | 0.160931714 | 0.689774323 |
| Residual | 58 | 0.311489964 | 0.005370517 | | |
| Total | 59 | 0.31235425 | | | |

| | *Coefficients* | *Standard Error* | *t-Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 0.00140958 | 0.01916082 | 0.073565754 | 0.941609047 | –0.036944967 | 0.039764128 |
| t | 0.000219156 | 0.000546301 | 0.401162952 | 0.689774323 | –0.000874385 | 0.001312696 |

**Figure 17-5:** Regression of Microsoft returns against time with a linear trend.

To run this regression, the independent variable (time) is assigned numerical values as follows. You assign the first date in the sample a value 1, the second date a value of 2, and so forth. So for this example, you assign August 2008 a value of 1, September 2008 a value of 2, and so on so that the last observation in the sample, July 2013, has a value of 60.

Note that in Figure 17-5, the coefficient of time is *not* statistically significant; its p-value is approximately 0.6898. For many hypothesis tests, as a rule of thumb any p-value above 0.05 indicates that a variable is not statistically significant.

More formally, the null hypothesis $H_0 : \beta_1 = 0$ can't be rejected at the 5 percent level of significance (see Chapter 12 for details on hypothesis testing.) This means there isn't enough evidence to show there is a trend in the data.

When there's no trend, the value of $y_t = \beta_0 + \varepsilon_t$.

As another example, suppose that instead of estimating a linear trend for the returns to Microsoft stock, you estimate a linear trend for the price of Microsoft stock. Figure 17-6 shows a plot of monthly Microsoft stock prices from August 2008 to July 2013.

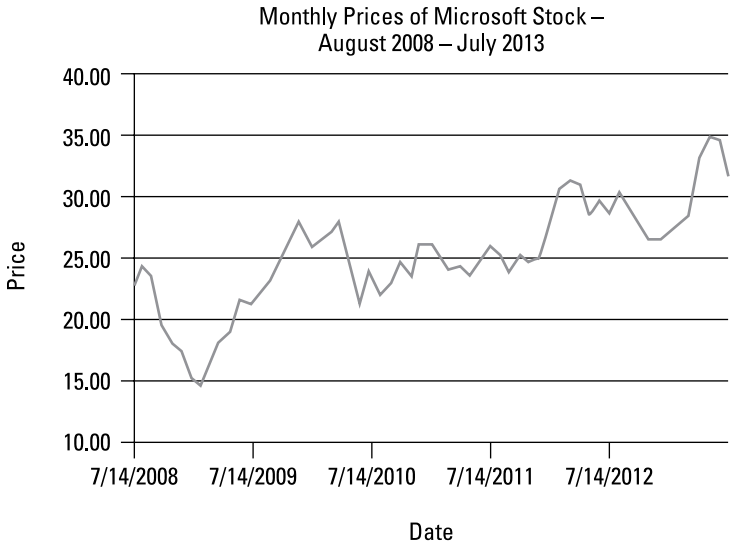



**Figure 17-6:** Monthly prices of Microsoft stock.

Figure 17-7 shows the results of running a regression of the price of Microsoft stock against time with an assumed linear trend.

The results show that the time variable is statistically significant at the 5 percent level (because the p-value for time is well below 0.05). Based on the coefficients in Figure 17-7, the estimated regression equation is $\hat{y}_t = 19.15 + 0.1975t$.

(Note that I rounded the coefficients in this equation.) This equation shows that during the sample period, the price of Microsoft stock grew by an average of $0.1975 per month because 0.1975 is the coefficient of *t,* and *y* is measured in dollars.

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.788249636 |
| R Square | 0.621337488 |
| Adjusted R Square | 0.614808824 |
| Standard Error | 2.715991268 |
| Observations | 60 |

ANOVA

| | *df* | *SS* | *MS* | *F-stat* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 702.0369613 | 702.0369613 | 95.17069461 | 7.7037E-14 |
| Residual | 58 | 427.843297 | 7.376608569 | | |
| Total | 59 | 1129.880258 | | | |

| | *Coefficients* | *Standard Error* | *t-Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 19.14990395 | 0.710124377 | 26.99697166 | 1.65632E-34 | 17.72843558 | 20.57137232 |
| t | 0.19751681 | 0.020246616 | 9.755546863 | 7.7037E-14 | 0.156988806 | 0.238044815 |

**Figure 17-7:** Regression of Microsoft prices against time with a linear trend.

Suppose that in your role as portfolio manager you want to determine whether a quadratic trend occurs in the time series of Microsoft stock prices.

If there is a quadratic trend in a time series, the appropriate regression equation is $y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$.

There is one new term in this equation:

$$\beta_2 t^2$$

Because time is squared here, this term captures the *curvature* of the trend. If this term is statistically significant, the trend associated with this time series is said to have a *quadratic* trend.

Figure 17-8 shows the results of running this regression.

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.78837024 |
| R Square | 0.621527635 |
| Adjusted R Square | 0.608247903 |
| Standard Error | 2.7390242 |
| Observations | 60 |

ANOVA

| | *df* | *SS* | *MS* | *F-stat* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 2 | 702.2518051 | 351.1259025 | 46.80272394 | 9.41833E-13 |
| Residual | 57 | 427.6284532 | 7.502253566 | | |
| Total | 59 | 1129.880258 | | | |

| | *Coefficients* | *Standard Error* | *t-Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 19.29057072 | 1.097189363 | 17.58180618 | 1.0881E-24 | 17.09348693 | 21.4876545 |
| t | 0.183903898 | 0.082993396 | 2.215886931 | 0.030707728 | 0.017712491 | 0.350095304 |
| $t^2$ | 0.000223163 | 0.00131873 | 0.169225377 | 0.8662185 | –0.002417548 | 0.002863873 |

**Figure 17-8:** Regression of Microsoft prices against time with a quadratic trend.

Figure 17-8 shows that the coefficient of time ($t$) is statistically significant, whereas the coefficient of time squared ($t^2$) is not, indicating that there is *not* a quadratic trend in the data, but there is a linear trend. Therefore, the price of Microsoft stock should be forecast with the linear trend model:

$$\hat{y}_t = 19.15 + 0.1975t$$

# Forecasting a Time Series

Based on the estimated regression equation from the previous section,

$$\hat{y}_t = 19.15 + 0.1975t$$

you can use this equation to predict the future value of Microsoft stock prices. By forecasting with a time series regression model, you are using the past history of Microsoft stock to make a prediction about where the stock will be in the future.

Suppose in July 2013 you want to forecast the price of Microsoft stock for August 2013. In the section "Estimating the Trend," the dates associated with the Microsoft stock prices are assigned numerical values ranging from 1 to 60; 60 represents the most recently observed price in July 2013. Therefore, August 2013 is assigned a value of 61. To forecast the price of Microsoft stock in August 2013, 61 is substituted for $t$ in the regression equation:

$$\hat{y}_t = 19.15 + 0.1975t$$
$$\hat{y}_t = 19.15 + 0.1975(61)$$
$$= \$31.1975$$
$$= (15.47,\ 39.17)$$

# Changing with the Seasons: Seasonal Variation

*Seasonal variation* refers to recurring changes in a time series that are due to the season of the year. For example, the demand for oil tends to be greatest during the summer (for gasoline) and the winter (for heating). For such cases, you extend the time series regression model to include a seasonal variable ($S_t$):

$$y_t = TR_t + S_t + \varepsilon_t$$

You then use a scatterplot to determine whether a time series exhibits seasonal variation, and if so, what type. For example, the seasonality could be quarterly or monthly.

To see the effect of seasonality, you can use *dummy variables.*

A dummy variable is also known as an indicator variable or a binary variable. A dummy variable is used to represent the values of a *qualitative* (non-numerical) variable in a regression equation; some examples are gender, color, style, and so on.

The most important feature of a dummy variable is that it can assume only one of two values: 1 or 0. 1 is normally used to indicate a specified condition is *true*, whereas 0 means that the condition is *false*. For example, a dummy variable could represent the gender of the people who reply to a survey. In this case, 1 could represent males and 0 could represent females.

For modeling seasonal variation, you can use a dummy variable to indicate whether an observation in a time series belongs to a given season. For example, suppose you're analyzing oil demand. You want to see whether the demand for oil is related to the quarter of the year. You have reason to believe that demand peaks in the fourth and first quarters due to cold weather.

For this exercise, you define the following seasonal dummy variables:

$D_1$ = 1 if time period $t$ is in the first quarter; it equals 0 otherwise.

$D_2$ = 1 if time period $t$ is in the second quarter; it equals 0 otherwise.

$D_3$ = 1 if time period $t$ is in the third quarter; it equals 0 otherwise.

In this case, you have only three dummy variables, not four, because including one dummy variable for each season leads to *multicollinearity* — when two or more independent variables in a regression equation are highly correlated with each other so they have large standard errors and can appear statistically insignificant even if they're not. Multicollinearity affects the reliability of the regression results, and can be avoided by not including highly correlated independent variables in the regression equation. (See Chapter 16 for more on multicollinearity.)

In this example, the coefficient of $D_1$ measures the impact on oil demand of the first quarter compared with the fourth quarter. In other words, the fourth quarter is used as a benchmark against which the other quarters are measured. If the coefficient of $D_1$ is positive, the demand for oil is *greater* in the first quarter than in the fourth quarter; if the coefficient of $D_1$ is negative, the demand for oil is *smaller* in the first quarter than in the fourth quarter. Similarly, the coefficient of $D_2$ measures the impact on oil demand of the second quarter compared with the fourth quarter, and the coefficient of $D_3$ measures the impact on oil demand of the third quarter compared with the fourth quarter.

The appropriate time series regression equation is

$$y_t = \beta_0 + \beta_1 t + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3 + \varepsilon_t$$

As an example, suppose a sporting goods store sells surfboards. In this case, sales depend heavily on the specific quarter of the year. In particular, sales are strongest during the second and third quarters and are extremely weak during the first and fourth quarters.

To analyze the relationship between surfboard sales and quarters, you run a regression with, say, ten years of quarterly data. Sales are the dependent variable. The independent variables consist of a time trend and a series of three quarterly dummy variables, defined as follows:

$D_1$ = 1 if an observation occurs in the first quarter, otherwise 0

$D_2$ = 1 if an observation occurs in the second quarter, otherwise 0

$D_3$ = 1 if an observation occurs in the third quarter, otherwise 0

The graph in Figure 17-9 shows quarterly sales (in thousands of dollars) of the sporting goods store for 2001 to 2010. A trend line is included.

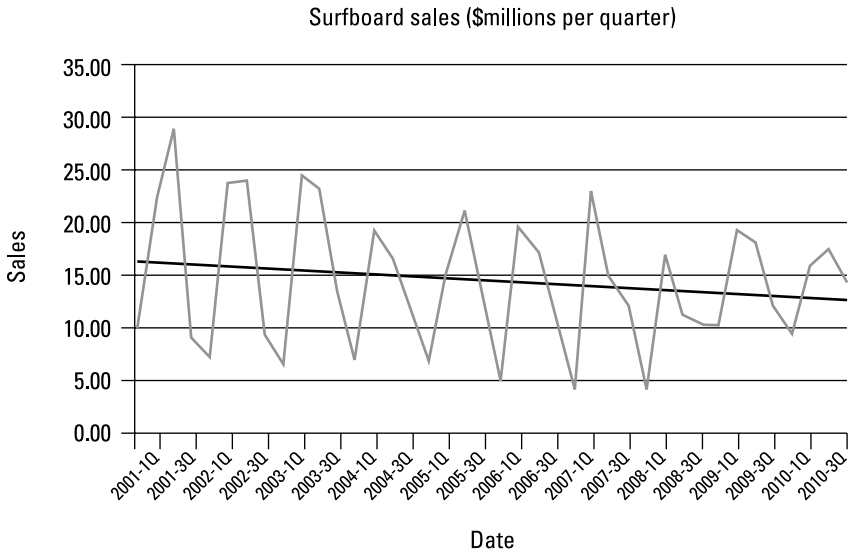



**Figure 17-9:** Quarterly sales data with seasonal variation.

Figure 17-9 shows that the trend line by itself does a poor job of explaining sales. The trend line is often extremely far from the actual sales numbers because the data are highly seasonal. Because the data are clearly affected by the seasons, it makes sense to run a regression with a trend and the seasonal dummies. Figure 17-10 shows the results.

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.883631708 |
| R Square | 0.780804995 |
| Adjusted R Square | 0.755754138 |
| Standard Error | 3.157746335 |
| Observations | 40 |

ANOVA

| | *df* | *SS* | *MS* | *F-stat* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 4 | 1243.181258 | 310.7953145 | 31.1687929 | 4.27358E-11 |
| Residual | 35 | 348.997667 | 9.971361915 | | |
| Total | 39 | 1592.178925 | | | |

| | *Coefficients* | *Standard Error* | *t-Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 13.90290102 | 1.382453507 | 10.05668614 | 7.31093E-12 | 11.09637121 | 16.70943083 |
| t | –0.10708641 | 0.043457054 | –2.46418938 | 0.018788462 | –0.195308919 | –0.018863901 |
| D1 | –4.855956296 | 1.418192167 | –3.42404676 | 0.001589404 | –7.73503944 | –1.976873152 |
| D2 | 8.246375747 | 1.414859161 | 5.828407502 | 1.29957E-06 | 5.374058965 | 11.11869253 |
| D3 | 7.657160168 | 1.412855583 | 5.419634009 | 4.49853E-06 | 4.788910866 | 10.52540947 |

**Figure 17-10:** Regression of quarterly sales data with seasonal dummies and trend.

The results show that each of the independent variables has a statistically significant coefficient and, therefore, belongs in the regression equation (in other words, these variables help explain the value of sales) because the p-value is below 0.05 in each case. Here are the approximate coefficients of the variables.

| | |
|---|---|
| ***Intercept*** | ***13.9029*** |
| Trend | –0.1071 |
| $D_1$ | –4.8560 |
| $D_2$ | 8.2464 |
| $D_3$ | 7.6572 |

The trend indicates that sales are decreasing by \$107.1 ($0.1071 \times \$1{,}000$) per month over the ten-year sample period. The coefficients of the remaining dummy variables show the value of sales compared with a trend line at the level of average fourth quarter sales. This line has an intercept of 13.9029 and a slope of –0.1071 and represents fourth quarter sales.

The coefficient of $D_1$ shows that sales during the first quarter are below the fourth quarter by \$4,856.00 ($4.8560 \times \$1{,}000$). The coefficient of $D_2$ shows that sales during the second quarter are above the fourth quarter by \$8,246.40 ($8.2464 \times 1{,}000$). The coefficient of $D_3$ shows that sales during the third quarter are above the fourth quarter by \$7,657.20 ($7.6572 \times \$1{,}000$).

# Implementing Smoothing Techniques

Smoothing techniques are designed to remove random fluctuations from a time series so the trend, seasonal variation, and cyclical variation (if any) in the data are easy to identify.

Two widely used smoothing techniques are *moving averages* and *centered moving averages,* which I talk about in the next sections.

## Moving averages

A *moving average* (MA) is an average of the most recent observations in a time series. For example, a five-period moving average is the average of the five most-recent values in a time series. In general, an *n-period moving average* is the average value of the *n* most recent observations taken from a time series.

Compute an *n*-period moving average with this formula:

$$\frac{y_t + y_{t+1} + y_{t+2} + \ldots + y_{t+n-1}}{n}$$

For example, the following lists the monthly prices of a stock between October 2012 and July 2013.

| ***Month*** | ***Price*** |
|---|---|
| October 2012 | 100 |
| November 2012 | 101 |
| December 2012 | 103 |
| January 2013 | 99 |
| February 2013 | 97 |
| March 2013 | 102 |
| April 2013 | 101 |
| May 2013 | 98 |
| June 2013 | 104 |
| July 2013 | 106 |

To construct a three-month moving average, take the average of the first three observations, in this case, October, November, and December prices:

$$\frac{(100+101+103)}{3} = 101.33$$

Then find the average of the next three observations, starting with the second observation, so you're finding the average of the second, third, and fourth observations (or November, December, and January):

$$\frac{(101+103+99)}{3} = 101.00$$

Continue the process for the entire sample. Table 17-1 shows the resulting three-month moving averages.

**Table 17-1 Three-Month Moving Averages**

| *Month* | *Price* | *3-Month Moving Average* |
|---|---|---|
| October 2012 | 100 | *** |
| November 2012 | 101 | 101.33 |
| December 2012 | 103 | 101.00 |
| January 2013 | 99 | 99.67 |
| February 2013 | 97 | 99.33 |
| March 2013 | 102 | 100.00 |
| April 2013 | 101 | 100.33 |
| May 2013 | 98 | 101.00 |
| June 2013 | 104 | 102.67 |
| July 2013 | 106 | *** |

The first three-month moving average is listed next to November 2012, even though it represents the average of October2, November2, and December2. This shows that November 2012 is the "center" of the moving average.

Similarly, the three-month moving average constructed from the November2, December2, and January3 prices is shown next to December, indicating that it's the center of the average2. Plotting these moving averages and the original prices (as shown in Figure 17-11) illustrates that moving averages reduce the fluctuations in the data and shows more clearly if there is any trend in the data. (The moving averages are said to "smooth out" the data.)

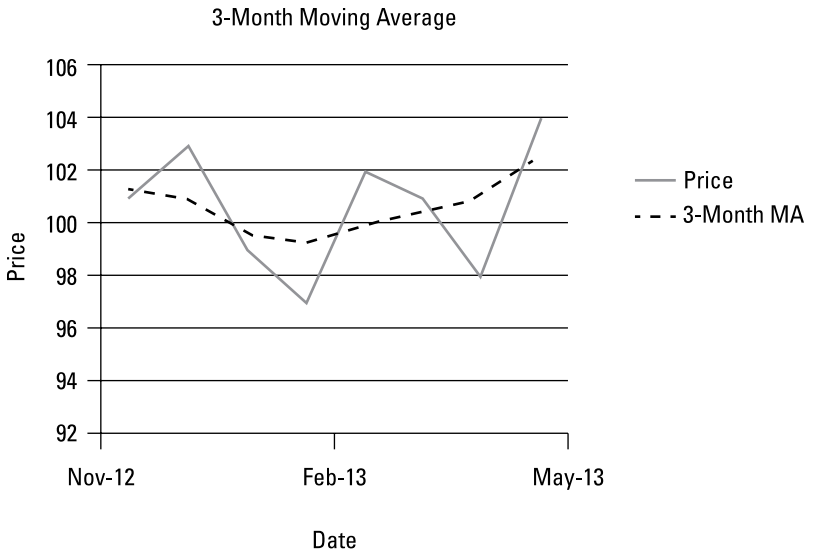



**Figure 17-11:** Original prices and three-month moving average.

The number of terms used to compute a moving average is usually determined by the data. For example, 12-month moving averages are often used with monthly data.

## Centered moving averages

A *centered moving average* is an average of moving averages. How's that for a definition? You use a centered moving average to remove the effect of seasonal and irregular factors from a time series, so only the trend and cyclical factors remain.

Using the stock prices from the previous example data (refer to Table 17-1), the first three-month moving average is 101.33 and the second three-month moving average is 101.

The centered moving average is then

$$\frac{(101.33+101)}{2}=101.17$$

Table 17-2 shows the centered moving averages for the rest of the months.

**Table 17-2 Three-Month Moving Averages and Centered Moving Averages**

| *Month* | *Price* | *3-Month Moving Average* | *Centered Moving Average* |
|---|---|---|---|
| October 2012 | 100 | *** | *** |
| November 2012 | 101 | 101.33 | 101.17 |
| December 2012 | 103 | 101.00 | 100.33 |
| January 2013 | 99 | 99.67 | 99.50 |
| February 2013 | 97 | 99.33 | 99.67 |
| March 2013 | 102 | 100.00 | 100.17 |
| April 2013 | 101 | 100.33 | 100.67 |
| May 2013 | 98 | 101.00 | 101.83 |
| June 2013 | 104 | 102.67 | *** |
| July 2013 | 106 | *** | *** |

Figure 17-12 shows a comparison of the moving average and centered moving average. The centered moving average is smoother than the moving average.

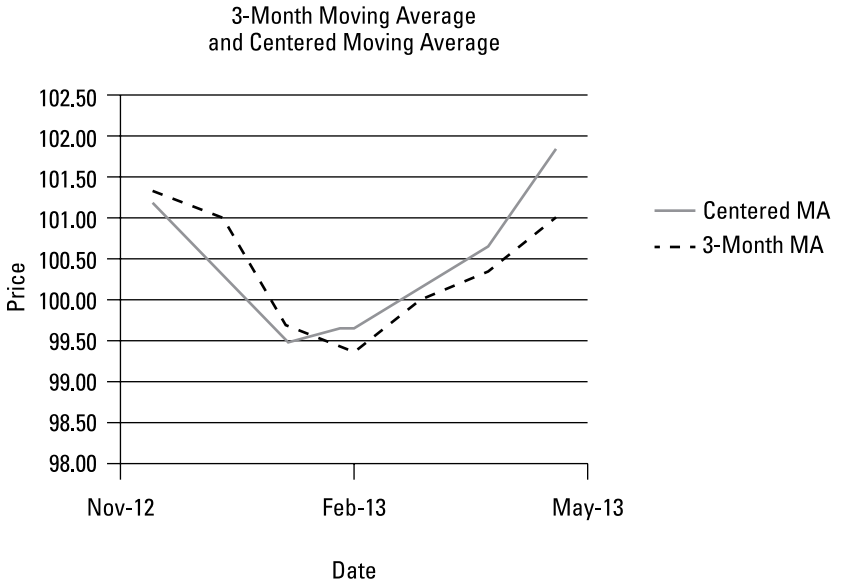



**Figure 17-12:** Prices, three-month moving averages, and centered moving averages.

# Exploring Exponential Smoothing

The moving average and centered moving average techniques have one common feature: Both assign equal weights to all elements of a time series. For example, when you're computing a three-month moving average, you multiply each observation by a weight of one-third (or, as you may know, you can get the same results by dividing by 3 instead). If a time series consists of data that become less relevant as time elapses, it may make more sense to assign steadily declining weights to older observations. You do this with *exponential smoothing*.

With exponential smoothing, you assign weights to the members of the time series to ensure that newer observations have more importance than older observations. You implement the weighting scheme using a *smoothing constant.* This is the value that determines how much smoothing takes place; the higher the smoothing constant, the more random variation is removed from the time series, thus making the time series smoother.

To implement the exponential smoothing approach, you use the following formula:

$$E_t = \alpha y_{t-1} + (1-\alpha) E_{t-1}$$

In this formula,

$E_t$ = the exponentially smoothed value at time $t$

$E_{t-1}$ = the exponentially smoothed value at time $t - 1$ (one period in the past)

$\alpha$ = the smoothing constant, which assumes a value between 0 and 1; the closer the value is to 1, the more smoothing takes place

$y_{t-1}$ = the value of the time series at time $t-1$

As an example, look at following lists of daily gold prices between 4/15/13 and 4/24/2013:

| ***Date*** | ***Price ($/ounce)*** |
|---|---|
| 4/15/13 | $1,481.84 |
| 4/16/13 | $1,422.82 |
| 4/17/13 | $1,368.21 |
| 4/18/13 | $1,378.20 |
| 4/19/13 | $1,381.07 |
| 4/20/13 | $1,401.96 |
| 4/21/13 | $1,403.53 |
| 4/22/13 | $1,403.53 |
| 4/23/13 | $1,421.14 |
| 4/24/13 | $1,418.78 |

An analyst wants to apply exponential smoothing to the data in order to produce a forecast of the price of gold on 4/25/13. Suppose the analyst believes that the data needs a significant amount of smoothing in order to eliminate random daily fluctuations in gold prices and show if there is any trend in the data. He picks a high value for the smoothing constant ($\alpha$); assume that he chooses 0.7 Table 17-3 shows the resulting exponentially smoothed values of the daily gold prices. (Assume that the exponentially smoothed price for 4/15/13 has already been computed from prior data to $1,493.77.)

**Table 17-3 Daily Gold Prices with Exponential Smoothing**

| *Date* | *Price ($/ounce)* | *Exponentially Smoothed Price ($\alpha$ = 0.7)* |
|---|---|---|
| 4/15/13 | $1,481.84 | $1493.77 |
| 4/16/13 | $1,422.82 | $1485.42 |
| 4/17/13 | $1,368.21 | $1441.60 |
| 4/18/13 | $1,378.20 | $1390.23 |
| 4/19/13 | $1,381.07 | $1381.81 |
| 4/20/13 | $1,401.96 | $1381.29 |
| 4/21/13 | $1,403.53 | $1395.76 |
| 4/22/13 | $1,403.53 | $1401.20 |
| 4/23/13 | $1,421.14 | $1402.83 |
| 4/24/13 | $1,418.78 | $1415.84 |

The exponentially smoothed price equals $\alpha$ times the previous day's price of gold plus $(1 - \alpha)$ times the previous day's exponentially smoothed price.

For example, on 4/16/13, the exponentially smoothed price is

$$E_t = \alpha y_{t-1} + (1-\alpha)E_{t-1}$$
$$E_t = (0.70)(1,481.84) + (0.30)(1,493.77)$$
$$= 1,485.42$$

On 4/17/13, the exponentially smoothed price is:

$$E_t = \alpha y_{t-1} + (1-\alpha)E_{t-1}$$
$$E_t = (0.70)(1,422.82) + (0.30)(1,485.42)$$
$$= 1,441.60$$

You compute the rest of the exponentially smoothed values the same way.

The graph in Figure 17-13 shows the relationship between actual gold prices and exponentially smoothed gold prices:

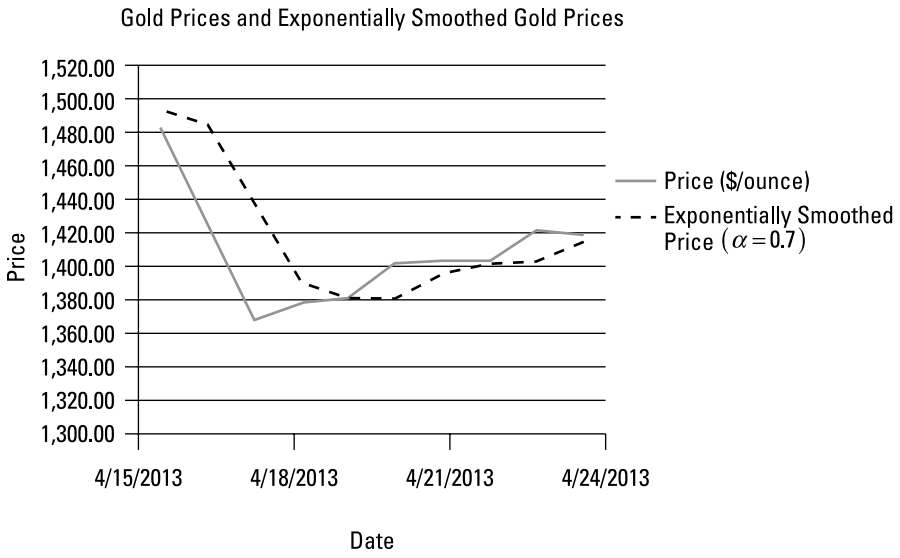



**Figure 17-13:** Prices and exponentially smoothed prices for gold.

As you can see, the exponentially weighted values don't fluctuate as much as the original values. With random fluctuations removed from the data, it is easier to see the trend in the data.

# Forecasting with exponential smoothing

With an exponential smoothing model, you can make a forecast for the next period with the following formula. The forecast for time $t + 1$ (one period in the future) as of time $t$ is $E_{t+1} = \alpha y_t + (1-\alpha)E_t$.

In the gold price example from the previous section, the price on 4/24/13 is $1,418.78, while the exponentially smoothed price is $1,415.84. The forecast for 4/25/13 is, therefore

$$E_{t+1} = (0.70)(1,418.78) + (0.30)(1,415.84)$$
$$= \$1,417.90$$

# Comparing the Forecasts of Different Models

Because there are several different types of models that can be used to predict the future values of a time series, it's important to be able to compare the quality of their results. Two techniques that are designed to test how well a forecasting model matches actual data are known as *mean absolute deviation (MAD)* and *mean square error (MSE).*

- **Mean absolute deviation (MAD)** is the average absolute value of the differences between the actual values of $y_t$ and the predicted values (for example, the absolute value of the *prediction errors*). You compute MAD with this formula:

$$MAD = \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t|}{n}$$

  $\hat{y}_t$ is the *predicted value* of $y_t$

  $y_t - \hat{y}_t$ is known as the *prediction error* associated with $y_t$

- **Mean square error (MSE)** is the average squared prediction error. You use the following equation to compute MSE:

$$MSE = \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n}$$

As an example, Figure 17-14 shows the prices of gold between 4/15/2013 and 4/24/2013. A time series model was used to forecast the price of gold during this period. A prediction error was computed for each date; the prediction error equals the actual price of gold minus the predicted value of gold. The absolute value of these prediction errors is computed for each date, as is the square of the prediction errors.

MAD is the average of the absolute values of the prediction errors; MSE is the average of the squared prediction errors. Figure 17-14 shows that the MAD equals 24.70, while the MSE equals 1079.44.

| DATE | PRICE | FORECAST | PREDICTION ERROR | ABSOLUTE VALUE OF PREDICTION ERROR | SQUARED PREDICTION ERROR |
|---|---|---|---|---|---|
| 4/15/2013 | 1,481.84 | 1,477.75 | 4.09 | 4.09 | 16.73 |
| 4/16/2013 | 1,422.82 | 1,451.57 | –28.75 | 28.75 | 826.36 |
| 4/17/2013 | 1,368.21 | 1,371.31 | –3.10 | 3.10 | 9.58 |
| 4/18/2013 | 1,378.20 | 1,383.23 | –5.03 | 5.03 | 25.28 |
| 4/19/2013 | 1,381.07 | 1,383.54 | –2.47 | 2.47 | 6.08 |
| 4/20/2013 | 1,401.96 | 1,461.27 | –59.31 | 59.31 | 3,518.21 |
| 4/21/2013 | 1,403.53 | 1,395.11 | 8.42 | 8.42 | 70.88 |
| 4/22/2013 | 1,403.53 | 1,455.87 | –52.34 | 52.34 | 2,739.84 |
| 4/23/2013 | 1,421.41 | 1,470.12 | –48.73 | 48.71 | 2,372.65 |
| 4/24/2013 | 1,418.78 | 1,384.01 | 34.77 | 34.77 | 1,208.80 |
| | | | **SUM** | **246.98** | **10,794.41** |
| | | | **AVERAGE** | **24.70** | **1,079.44** |

**Figure 17-14:** MAD and MSE computed for gold price forecasts.

For any type of predictive model, the lower the value of the MAD or the MSE, the better the model fits the observed data. Using these measures lets you compare the results of different models (such as moving averages, exponential smoothing, and so forth) to determine which model provides the most accurate predictions for a given set of data.

One of the drawbacks of MSE is that it's more affected by extremely large prediction errors than MAD. One of the advantages of MSE is that it has more convenient mathematical properties than MAD. Because MAD is based on the absolute value, techniques for minimizing MAD are more complex than techniques for minimizing MSE.