CHAPTER 12

Simple Regression Analysis and Correlation

LEARNING OBJECTIVES

The overall objective of this chapter is to give you an understanding of bivariate linear regression analysis, thereby enabling you to:

1. Calculate the Pearson product-moment correlation coefficient to determine if there is a correlation between two variables.
2. Explain what regression analysis is and the concepts of independent and dependent variable.
3. Calculate the slope and y-intercept of the least squares equation of a regression line and from those, determine the equation of the regression line.
4. Calculate the residuals of a regression line and from those determine the fit of the model, locate outliers, and test the assumptions of the regression model.
5. Calculate the standard error of the estimate using the sum of squares of error, and use the standard error of the estimate to determine the fit of the model.
6. Calculate the coefficient of determination to measure the fit for regression models, and relate it to the coefficient of correlation.
7. Use the *t* and *F* tests to test hypotheses for both the slope of the regression model and the overall regression model.
8. Calculate confidence intervals to estimate the conditional mean of the dependent variable and prediction intervals to estimate a single value of the dependent variable.
9. Determine the equation of the trend line to forecast outcomes for time periods in the future, using alternate coding for time periods if necessary.
10. Use a computer to develop a regression analysis, and interpret the output that is associated with it.

Najlah Feanny/©Corbis

Predicting International Hourly Wages by the Price of a Big Mac

Decision Dilemma

The McDonald's Corporation is the leading global foodservice retailer with more than 30,000 local restaurants serving nearly 50 million people in more than 119 countries each day. This global presence, in addition to its consistency in food offerings and restaurant operations, makes McDonald's a unique and attractive setting for economists to make salary and price comparisons around the world. Because the Big Mac hamburger is a standardized hamburger produced and sold in virtually every McDonald's around the world, the *Economist,* a weekly newspaper focusing on international politics and business news and opinion, as early as 1986 was compiling information about Big Mac prices as an indicator of exchange rates. Building on this idea, researchers Ashenfelter and Jurajda proposed comparing wage rates across countries and the price of a Big Mac hamburger. Shown below are Big Mac prices and net hourly wage figures (in U.S. dollars) for 27 countries. Note that net hourly wages are based on a weighted average of 12 professions.

Country	Big Mac Price (U.S. $)	Net Hourly Wage (U.S. $)
Argentina	1.42	1.70
Australia	1.86	7.80
Brazil	1.48	2.05
Britain	3.14	12.30
Canada	2.21	9.35
Chile	1.96	2.80
China	1.20	2.40
Czech Republic	1.96	2.40
Denmark	4.09	14.40
Euro area	2.98	9.59
Hungary	2.19	3.00
Indonesia	1.84	1.50
Japan	2.18	13.60
Malaysia	1.33	3.10
Mexico	2.18	2.00
New Zealand	2.22	6.80
Philippines	2.24	1.20
Poland	1.62	2.20
Russia	1.32	2.60
Singapore	1.85	5.40
South Africa	1.85	3.90
South Korea	2.70	5.90
Sweden	3.60	10.90
Switzerland	4.60	17.80
Thailand	1.38	1.70
Turkey	2.34	3.20
United States	2.71	14.30

(*continued*)

Managerial and Statistical Questions

1. Is there a relationship between the price of a Big Mac and the net hourly wages of workers around the world? If so, how strong is the relationship?
2. Is it possible to develop a model to predict or determine the net hourly wage of a worker around the world by the price of a Big Mac hamburger in that country? If so, how good is the model?
3. If a model can be constructed to determine the net hourly wage of a worker around the world by the price of a Big Mac hamburger, what would be the predicted net hourly wage of a worker in a country if the price of a Big Mac hamburger was $3.00?

Sources: McDonald's Web site at: http://www.mcdonalds.com/corp/about.html; Michael R. Pakko and Patricia S. Pollard, "Burgernomics: A Big Mac Guide to Purchasing Power Parity," research publication by the St. Louis Federal Reserve Bank at: http://research.stlouisfed.org/publications/review/03/11/pakko.pdf; Orley Ashenfelter and Stěpán Jurajda, "Cross-Country Comparisons of Wage Rates: The Big Mac Index," unpublished manuscript, Princeton University and CERGEEI/Charles University, October 2001; *The Economist,* at: http://www.economist.com/index.html.

In business, the key to decision making often lies in the understanding of the relationships between two or more variables. For example, a company in the distribution business may determine that there is a relationship between the price of crude oil and their own transportation costs. Financial experts, in studying the behavior of the bond market, might find it useful to know if the interest rates on bonds are related to the prime

interest rate set by the Federal Reserve. A marketing executive might want to know how strong the relationship is between advertising dollars and sales dollars for a product or a company.

In this chapter, we will study the concept of correlation and how it can be used to estimate the relationship between two variables. We will also explore simple regression analysis through which mathematical models can be developed to predict one variable by another. We will examine tools for testing the strength and predictability of regression models, and we will learn how to use regression analysis to develop a forecasting trend line.

12.1 CORRELATION

Correlation is *a measure of the degree of relatedness of variables.* It can help a business researcher determine, for example, whether the stocks of two airlines rise and fall in any related manner. For a sample of pairs of data, correlation analysis can yield a numerical value that represents the degree of relatedness of the two stock prices over time. In the transportation industry, is a correlation evident between the price of transportation and the weight of the object being shipped? If so, how strong are the correlations? In economics, how strong is the correlation between the producer price index and the unemployment rate? In retail sales, are sales related to population density, number of competitors, size of the store, amount of advertising, or other variables?

TABLE 12.1
Data for the Economics Example

Day	Interest Rate	Futures Index
1	7.43	221
2	7.48	222
3	8.00	226
4	7.75	225
5	7.60	224
6	7.63	223
7	7.68	223
8	7.67	226
9	7.59	226
10	8.07	235
11	8.03	233
12	8.00	241

Several measures of correlation are available, the selection of which depends mostly on the level of data being analyzed. Ideally, researchers would like to solve for ρ, the population coefficient of correlation. However, because researchers virtually always deal with sample data, this section introduces a widely used sample **coefficient of correlation**, r. This measure is applicable only if both variables being analyzed have at least an interval level of data. Chapter 17 presents a correlation measure that can be used when the data are ordinal.

The statistic r is the **Pearson product-moment correlation coefficient**, named after Karl Pearson (1857–1936), an English statistician who developed several coefficients of correlation along with other significant statistical concepts. The term r is a *measure of the linear correlation of two variables.* It is a number that ranges from −1 to 0 to +1, representing the strength of the relationship between the variables. An r value of +1 denotes a perfect positive relationship between two sets of numbers. An r value of −1 denotes a perfect negative correlation, which indicates an inverse relationship between two variables: as one variable gets larger, the other gets smaller. An r value of 0 means no linear relationship is present between the two variables.

PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT (12.1)

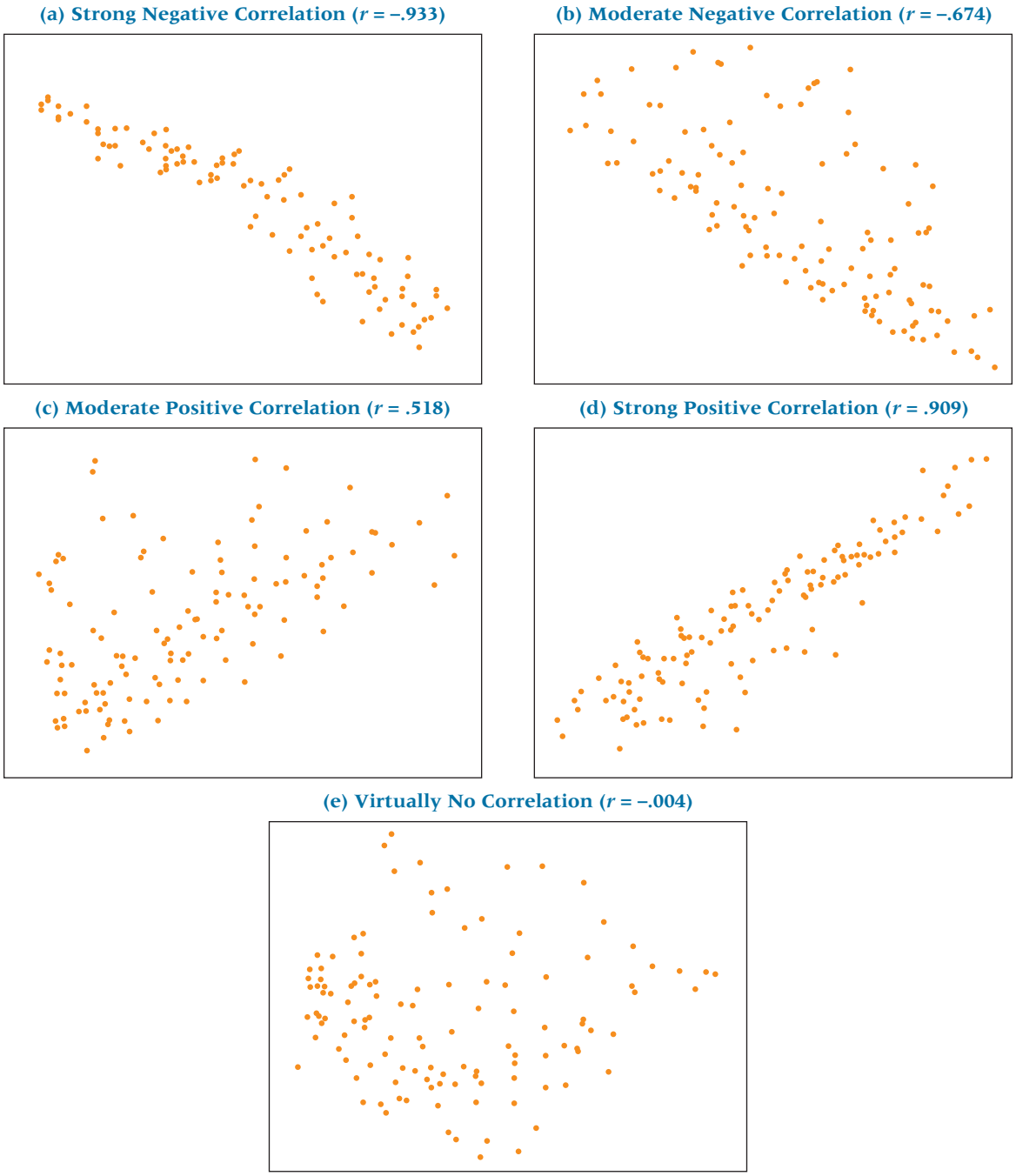
$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}} = \frac{\Sigma xy - \dfrac{(\Sigma x\Sigma y)}{n}}{\sqrt{\left[\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}\right]}}$$

Figure 12.1 depicts five different degrees of correlation: (a) represents strong negative correlation, (b) represents moderate negative correlation, (c) represents moderate positive correlation, (d) represents strong positive correlation, and (e) contains no correlation.

What is the measure of correlation between the interest rate of federal funds and the commodities futures index? With data such as those shown in Table 12.1, which represent the values for interest rates of federal funds and commodities futures indexes for a sample of 12 days, a correlation coefficient, r, can be computed.

FIGURE 12.1

Five Correlations



Examination of the formula for computing a Pearson product-moment correlation coefficient (12.1) reveals that the following values must be obtained to compute r: Σx, Σx^2, Σy, Σy^2, Σxy, and n. In correlation analysis, it does not matter which variable is designated x and which is designated y. For this example, the correlation coefficient is computed as shown in Table 12.2. The r value obtained ($r = .815$) represents a relatively strong positive relationship between interest rates and commodities futures index over this 12-day period.

Figure 12.2 shows both Excel and Minitab output for this problem.

TABLE 12.2
Computation of r for the Economics Example

Day	Interest x	Futures Index y	x^2	y^2	xy
1	7.43	221	55.205	48,841	1,642.03
2	7.48	222	55.950	49,284	1,660.56
3	8.00	226	64.000	51,076	1,808.00
4	7.75	225	60.063	50,625	1,743.75
5	7.60	224	57.760	50,176	1,702.40
6	7.63	223	58.217	49,729	1,701.49
7	7.68	223	58.982	49,729	1,712.64
8	7.67	226	58.829	51,076	1,733.42
9	7.59	226	57.608	51,076	1,715.34
10	8.07	235	65.125	55,225	1,896.45
11	8.03	233	64.481	54,289	1,870.99
12	8.00	241	64.000	58,081	1,928.00
	$\Sigma x = 92.93$	$\Sigma y = 2{,}725$	$\Sigma x^2 = 720.220$	$\Sigma y^2 = 619{,}207$	$\Sigma xy = 21{,}115.07$

$$r = \frac{(21{,}115.07) - \dfrac{(92.93)(2725)}{12}}{\sqrt{\left[(720.22) - \dfrac{(92.93)^2}{12}\right]\left[(619{,}207) - \dfrac{(2725)^2}{12}\right]}} = .815$$

FIGURE 12.2
Excel and Minitab Output for the Economics Example

Excel Output

	Interest Rate	*Futures Index*
Interest Rate	1	
Futures Index	0.815	1

Minitab Output

Correlations: Interest Rate, Futures Index

```
Pearson correlation of Interest Rate and Futures Index = 0.815
p-Value = 0.001
```

12.1 PROBLEMS

12.1 Determine the value of the coefficient of correlation, r, for the following data.

X	4	6	7	11	14	17	21
Y	18	12	13	8	7	7	4

12.2 Determine the value of r for the following data.

X	158	296	87	110	436
Y	349	510	301	322	550

12.3 In an effort to determine whether any correlation exists between the price of stocks of airlines, an analyst sampled six days of activity of the stock market. Using the following prices of Delta stock and Southwest stock, compute the coefficient of correlation. Stock prices have been rounded off to the nearest tenth for ease of computation.

Delta	Southwest
47.6	15.1
46.3	15.4
50.6	15.9
52.6	15.6
52.4	16.4
52.7	18.1

12.4 The following data are the claims (in $ millions) for BlueCross BlueShield benefits for nine states, along with the surplus (in $ millions) that the company had in assets in those states.

State	Claims	Surplus
Alabama	$1,425	$277
Colorado	273	100
Florida	915	120
Illinois	1,687	259
Maine	234	40
Montana	142	25
North Dakota	259	57
Oklahoma	258	31
Texas	894	141

Use the data to compute a correlation coefficient, *r*, to determine the correlation between claims and surplus.

12.5 The National Safety Council released the following data on the incidence rates for fatal or lost-worktime injuries per 100 employees for several industries in three recent years.

Industry	Year 1	Year 2	Year 3
Textile	.46	.48	.69
Chemical	.52	.62	.63
Communication	.90	.72	.81
Machinery	1.50	1.74	2.10
Services	2.89	2.03	2.46
Nonferrous metals	1.80	1.92	2.00
Food	3.29	3.18	3.17
Government	5.73	4.43	4.00

Compute *r* for each pair of years and determine which years are most highly correlated.

12.2 INTRODUCTION TO SIMPLE REGRESSION ANALYSIS

Regression analysis is *the process of constructing a mathematical model or function that can be used to predict or determine one variable by another variable or other variables.* The most elementary regression model is called **simple regression** or **bivariate regression** involving two variables in which one variable is predicted by another variable. In simple regression, *the variable to be predicted* is called the **dependent variable** and is designated as *y*. The *predictor* is called the **independent variable**, or *explanatory variable*, and is designated as *x*. In simple regression analysis, only a straight-line relationship between two variables is examined. Nonlinear relationships and regression models with more than one independent variable can be explored by using multiple regression models, which are presented in Chapters 13 and 14.

TABLE 12.3
Airline Cost Data

Number of Passengers	Cost ($1,000)
61	4.280
63	4.080
67	4.420
69	4.170
70	4.480
74	4.300
76	4.820
81	4.700
86	5.110
91	5.130
95	5.640
97	5.560

Can the cost of flying a commercial airliner be predicted using regression analysis? If so, what variables are related to such cost? A few of the many variables that can potentially contribute are type of plane, distance, number of passengers, amount of luggage/freight, weather conditions, direction of destination, and perhaps even pilot skill. Suppose a study is conducted using only Boeing 737s traveling 500 miles on comparable routes during the same season of the year. Can the number of passengers predict the cost of flying such routes? It seems logical that more passengers result in more weight and more baggage, which could, in turn, result in increased fuel consumption and other costs. Suppose the data displayed in Table 12.3 are the costs and associated number of passengers for twelve 500-mile commercial airline flights using Boeing 737s during the same season of the year. We will use these data to develop a regression model to predict cost by number of passengers.

Usually, the first step in simple regression analysis is to construct a **scatter plot** (or scatter diagram), discussed in Chapter 2. Graphing the data in this way yields preliminary information about the shape and spread of the data. Figure 12.3 is an Excel scatter plot of the data in Table 12.3. Figure 12.4 is a close-up view of the scatter plot produced by

FIGURE 12.3
Excel Scatter Plot of Airline Cost Data

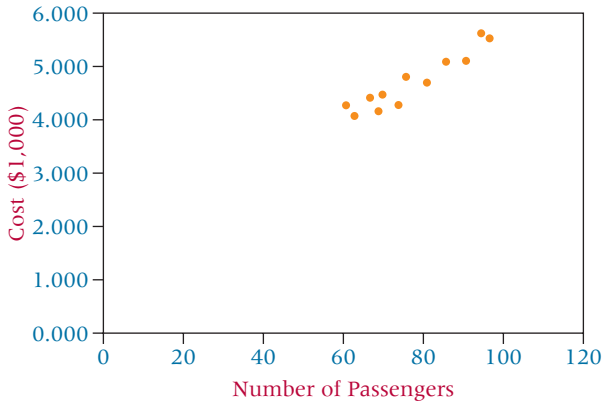
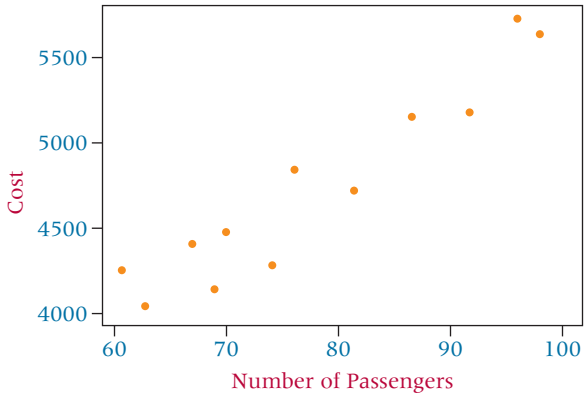


FIGURE 12.4
Close-Up Minitab Scatter Plot of Airline Cost Data



Minitab. Try to imagine a line passing through the points. Is a linear fit possible? Would a curve fit the data better? The scatter plot gives some idea of how well a regression line fits the data. Later in the chapter, we present statistical techniques that can be used to determine more precisely how well a regression line fits the data.

12.3 DETERMINING THE EQUATION OF THE REGRESSION LINE

Interactive Applet

The first step in determining the equation of the regression line that passes through the sample data is to establish the equation's form. Several different types of equations of lines are discussed in algebra, finite math, or analytic geometry courses. Recall that among these equations of a line are the two-point form, the point-slope form, and the slope-intercept form. In regression analysis, researchers use the slope-intercept equation of a line. In math courses, the slope-intercept form of the equation of a line often takes the form

$$y = mx + b$$

where

m = slope of the line
b = y intercept of the line

In statistics, the slope-intercept form of the equation of the regression line through the population points is

$$\hat{y} = \beta_0 + \beta_1 x$$

where

$\hat{y}$ = the predicted value of y
β_0 = the population y intercept
β_1 = the population slope

For any specific dependent variable value, y_i,

$$y_i = \beta_0 + \beta_1 x_i + \in_i$$

where

x_i = the value of the independent variable for the ith value
y_i = the value of the dependent variable for the ith value
β_0 = the population y intercept
β_1 = the population slope
$\in_i$ = the error of prediction for the ith value

Unless the points being fitted by the regression equation are in perfect alignment, the regression line will miss at least some of the points. In the preceding equation, $\in_i$ represents the error of the regression line in fitting these points. If a point is on the regression line, $\in_i = 0$.

These mathematical models can be either deterministic models or probabilistic models. **Deterministic models** are *mathematical models that produce an "exact" output for a given input.* For example, suppose the equation of a regression line is

$$y = 1.68 + 2.40x$$

For a value of $x = 5$, the exact predicted value of y is

$$y = 1.68 + 2.40(5) = 13.68$$

We recognize, however, that most of the time the values of y will not equal exactly the values yielded by the equation. Random error will occur in the prediction of the y values for values of x because it is likely that the variable x does not explain all the variability of the variable y. For example, suppose we are trying to predict the volume of sales (y) for a company through regression analysis by using the annual dollar amount of advertising (x) as the predictor. Although sales are often related to advertising, other factors related to sales are not accounted for by amount of advertising. Hence, a regression model to predict sales volume by amount of advertising probably involves some error. For this reason, in regression, we present the general model as a probabilistic model. A **probabilistic model** is *one that includes an error term that allows for the y values to vary for any given value of x.*

A deterministic regression model is

$$y = \beta_0 + \beta_1 x$$

The probabilistic regression model is

$$y = \beta_0 + \beta_1 x + \in$$

$\beta_0 + \beta_1 x$ is the deterministic portion of the probabilistic model, $\beta_0 + \beta_1 x + \in$. In a deterministic model, all points are assumed to be on the line and in all cases $\in$ is zero.

Virtually all regression analyses of business data involve sample data, not population data. As a result, β_0 and β_1 are unattainable and must be estimated by using the sample statistics, b_0 and b_1. Hence the equation of the regression line contains the sample y intercept, b_0, and the sample slope, b_1.

EQUATION OF THE SIMPLE REGRESSION LINE

$$\hat{y} = b_0 + b_1 x$$

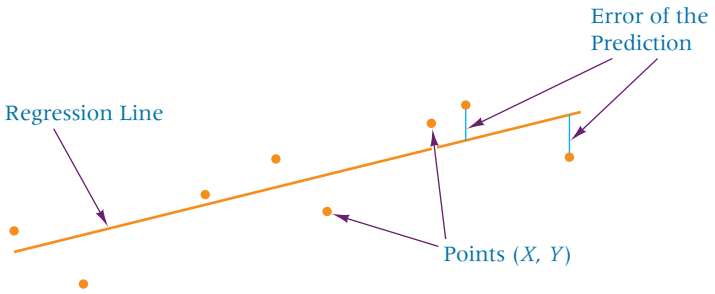
Where

b_0 = the sample intercept
b_1 = the sample slope

To determine the equation of the regression line for a sample of data, the researcher must determine the values for b_0 and b_1. This process is sometimes referred to as least squares analysis. **Least squares analysis** is *a process whereby a regression model is developed by producing the minimum sum of the squared error values.* On the basis of this premise and calculus, a particular set of equations has been developed to produce components of the regression model.*

*Derivation of these formulas is beyond the scope of information being discussed here but is presented in WileyPLUS.

FIGURE 12.5
Minitab Plot of a Regression Line



Examine the regression line fit through the points in Figure 12.5. Observe that the line does not actually pass through any of the points. The vertical distance from each point to the line is the error of the prediction. In theory, an infinite number of lines could be constructed to pass through these points in some manner. The least squares regression line is the regression line that results in the smallest sum of errors squared.

Formula 12.2 is an equation for computing the value of the sample slope. Several versions of the equation are given to afford latitude in doing the computations.

SLOPE OF THE REGRESSION LINE (12.2)

$$b_1 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2} = \frac{\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}$$

The expression in the numerator of the slope formula 12.2 appears frequently in this chapter and is denoted as SS_{xy}.

$$SS_{xy} = \Sigma(x - \bar{x})(y - \bar{y}) = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$$

The expression in the denominator of the slope formula 12.2 also appears frequently in this chapter and is denoted as SS_{xx}.

$$SS_{xx} = \Sigma(x - \bar{x})^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

With these abbreviations, the equation for the slope can be expressed as in Formula 12.3.

ALTERNATIVE FORMULA FOR SLOPE (12.3)

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

Formula 12.4 is used to compute the sample y intercept. The slope must be computed before the y intercept.

y INTERCEPT OF THE REGRESSION LINE (12.4)

$$b_0 = \bar{y} - b_1\bar{x} = \frac{\Sigma y}{n} - b_1\frac{(\Sigma x)}{n}$$

Formulas 12.2, 12.3, and 12.4 show that the following data are needed from sample information to compute the slope and intercept: Σx, Σy, Σx^2, and, Σxy, unless sample means are used. Table 12.4 contains the results of solving for the slope and intercept and determining the equation of the regression line for the data in Table 12.3.

The least squares equation of the regression line for this problem is

$$\hat{y} = 1.57 + .0407x$$

TABLE 12.4

Solving for the Slope and the y Intercept of the Regression Line for the Airline Cost Example

Number of Passengers	Cost ($1,000)		
x	y	x^2	xy
61	4.280	3,721	261.080
63	4.080	3,969	257.040
67	4.420	4,489	296.140
69	4.170	4,761	287.730
70	4.480	4,900	313.600
74	4.300	5,476	318.200
76	4.820	5,776	366.320
81	4.700	6,561	380.700
86	5.110	7,396	439.460
91	5.130	8,281	466.830
95	5.640	9,025	535.800
97	5.560	9,409	539.320
$\Sigma x = 930$	$\Sigma y = 56.690$	$\Sigma x^2 = 73,764$	$\Sigma xy = 4462.220$

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 4462.22 - \frac{(930)(56.69)}{12} = 68.745$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x^2)}{n} = 73,764 - \frac{(930)^2}{12} = 1689$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{68.745}{1689} = .0407$$

$$b_0 = \frac{\Sigma y}{n} - b_1\frac{\Sigma x}{n} = \frac{56.19}{12} - (.0407)\frac{930}{12} = 1.57$$

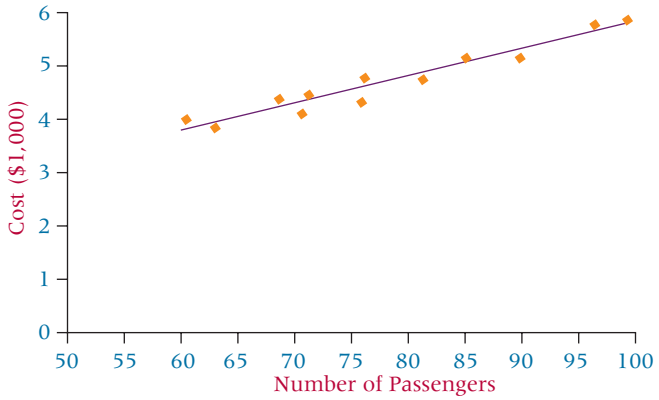
$$\hat{y} = 1.57 + .0407x$$

The slope of this regression line is .0407. Because the x values were recoded for the ease of computation and are actually in $1,000 denominations, the slope is actually $40.70. One interpretation of the slope in this problem is that for every unit increase in x (every person added to the flight of the airplane), there is a $40.70 increase in the cost of the flight. The y-intercept is the point where the line crosses the y-axis (where x is zero). Sometimes in regression analysis, the y-intercept is meaningless in terms of the variables studied. However, in this problem, one interpretation of the y-intercept, which is 1.570 or $1,570, is that even if there were no passengers on the commercial flight, it would still cost $1,570. In other words, there are costs associated with a flight that carries no passengers.

Superimposing the line representing the least squares equation for this problem on the scatter plot indicates how well the regression line fits the data points, as shown in the Excel graph in Figure 12.6. The next several sections explore mathematical ways of testing how well the regression line fits the points.

FIGURE 12.6

Excel Graph of Regression Line for the Airline Cost Example



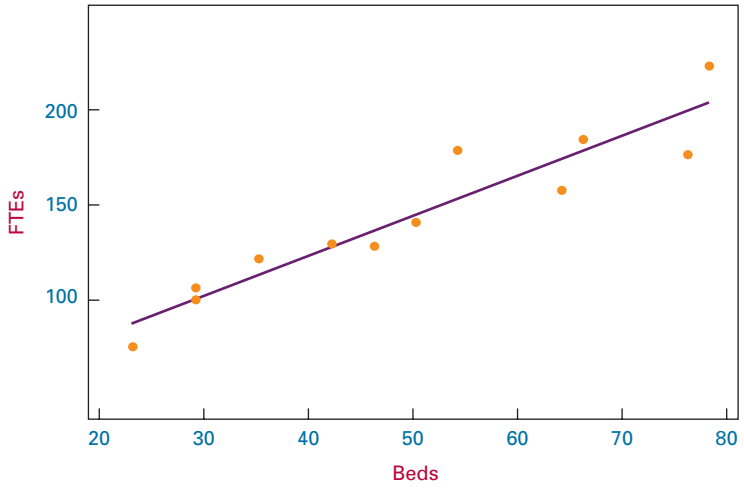
DEMONSTRATION PROBLEM 12.1

A specialist in hospital administration stated that the number of FTEs (full-time employees) in a hospital can be estimated by counting the number of beds in the hospital (a common measure of hospital size). A healthcare business researcher decided to develop a regression model in an attempt to predict the number of FTEs of a hospital by the number of beds. She surveyed 12 hospitals and obtained the following data. The data are presented in sequence, according to the number of beds.

Number of Beds	FTEs	Number of Beds	FTEs
23	69	50	138
29	95	54	178
29	102	64	156
35	118	66	184
42	126	76	176
46	125	78	225

Solution

The following Minitab graph is a scatter plot of these data. Note the linear appearance of the data.



Next, the researcher determined the values of Σx, Σy, Σx^2, and Σxy.

Hospital	Number of Beds x	FTEs y	x^2	xy
1	23	69	529	1,587
2	29	95	841	2,755
3	29	102	841	2,958
4	35	118	1,225	4,130
5	42	126	1,764	5,292
6	46	125	2,116	5,750
7	50	138	2,500	6,900
8	54	178	2,916	9,612
9	64	156	4,096	9,984
10	66	184	4,356	12,144
11	76	176	5,776	13,376
12	78	225	6,084	17,550
	$\Sigma x = 592$	$\Sigma y = 1{,}692$	$\Sigma x^2 = 33{,}044$	$\Sigma xy = 92{,}038$

Using these values, the researcher solved for the sample slope (b_1) and the sample y-intercept (b_0).

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 92{,}038 - \frac{(592)(1692)}{12} = 8566$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 33{,}044 - \frac{(592)^2}{12} = 3838.667$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{8566}{3838.667} = 2.232$$

$$b_0 = \frac{\Sigma y}{n} - b_1\frac{\Sigma x}{12} = \frac{1692}{12} - (2.232)\frac{592}{12} = 30.888$$

The least squares equation of the regression line is

$$\hat{y} = 30.888 + 2.232x$$

The slope of the line, $b_1 = 2.232$, means that for every unit increase of x (every bed), y (number of FTEs) is predicted to increase by 2.232. Even though the y-intercept helps the researcher sketch the graph of the line by being one of the points on the line (0, 30.888), it has limited usefulness in terms of this solution because $x = 0$ denotes a hospital with no beds. On the other hand, it could be interpreted that a hospital has to have at least 31 FTEs to open its doors even with no patients—a sort of "fixed cost" of personnel.

12.3 PROBLEMS

12.6 Sketch a scatter plot from the following data, and determine the equation of the regression line.

x	12	21	28	8	20
y	17	15	22	19	24

12.7 Sketch a scatter plot from the following data, and determine the equation of the regression line.

x	140	119	103	91	65	29	24
y	25	29	46	70	88	112	128

12.8 A corporation owns several companies. The strategic planner for the corporation believes dollars spent on advertising can to some extent be a predictor of total sales dollars. As an aid in long-term planning, she gathers the following sales and advertising information from several of the companies for 2009 ($ millions).

Advertising	Sales
12.5	148
3.7	55
21.6	338
60.0	994
37.6	541
6.1	89
16.8	126
41.2	379

Develop the equation of the simple regression line to predict sales from advertising expenditures using these data.

12.9 Investment analysts generally believe the interest rate on bonds is inversely related to the prime interest rate for loans; that is, bonds perform well when lending rates are down and perform poorly when interest rates are up. Can the bond rate be predicted by the prime interest rate? Use the following data to construct a least squares regression line to predict bond rates by the prime interest rate.

Bond Rate	Prime Interest Rate
5%	16%
12	6
9	8
15	4
7	7

12.10 Is it possible to predict the annual number of business bankruptcies by the number of firm births (business starts) in the United States? The following data published by the U.S. Small Business Administration, Office of Advocacy, are pairs of the number of business bankruptcies (1000s) and the number of firm births (10,000s) for a six-year period. Use these data to develop the equation of the regression model to predict the number of business bankruptcies by the number of firm births. Discuss the meaning of the slope.

Business Bankruptcies (1000)	Firm Births (10,000)
34.3	58.1
35.0	55.4
38.5	57.0
40.1	58.5
35.5	57.4
37.9	58.0

12.11 It appears that over the past 45 years, the number of farms in the United States declined while the average size of farms increased. The following data provided by the U.S. Department of Agriculture show five-year interval data for U.S. farms. Use these data to develop the equation of a regression line to predict the average size of a farm by the number of farms. Discuss the slope and y-intercept of the model.

Year	Number of Farms (millions)	Average Size (acres)
1950	5.65	213
1955	4.65	258
1960	3.96	297
1965	3.36	340
1970	2.95	374
1975	2.52	420
1980	2.44	426
1985	2.29	441
1990	2.15	460
1995	2.07	469
2000	2.17	434
2005	2.10	444

12.12 Can the annual new orders for manufacturing in the United States be predicted by the raw steel production in the United States? Shown on the next page are the annual new orders for 10 years according to the U.S. Census Bureau and the raw steel production for the same 10 years as published by the American Iron & Steel Institute. Use these data to develop a regression model to predict annual new orders by raw steel production. Construct a scatter plot and draw the regression line through the points.

Raw Steel Production (100,000s of net tons)	New Orders ($ trillions)
99.9	2.74
97.9	2.87
98.9	2.93
87.9	2.87
92.9	2.98
97.9	3.09
100.6	3.36
104.9	3.61
105.3	3.75
108.6	3.95

12.4 RESIDUAL ANALYSIS

How does a business researcher test a regression line to determine whether the line is a good fit of the data other than by observing the fitted line plot (regression line fit through a scatter plot of the data)? One particularly popular approach is to use the *historical data* (x and y values used to construct the regression model) to test the model. With this approach, the values of the independent variable (x values) are inserted into the regression model and a predicted value ($\hat{y}$) is obtained for each x value. These predicted values ($\hat{y}$) are then compared to the actual y values to determine how much error the equation of the regression line produced. *Each difference between the actual y values and the predicted y values is the error of the regression line at a given point,* $y - \hat{y}$, and is referred to as the **residual**. It is the sum of squares of these residuals that is minimized to find the least squares line.

Table 12.5 shows $\hat{y}$ values and the residuals for each pair of data for the airline cost regression model developed in Section 12.3. The predicted values are calculated by inserting an x value into the equation of the regression line and solving for $\hat{y}$. For example, when $x = 61$, $\hat{y} = 1.57 + .0407(61) = 4.053$, as displayed in column 3 of the table. Each of these predicted y values is subtracted from the actual y value to determine the error, or residual. For example, the first y value listed in the table is 4.280 and the first predicted value is 4.053, resulting in a residual of $4.280 - 4.053 = .227$. The residuals for this problem are given in column 4 of the table.

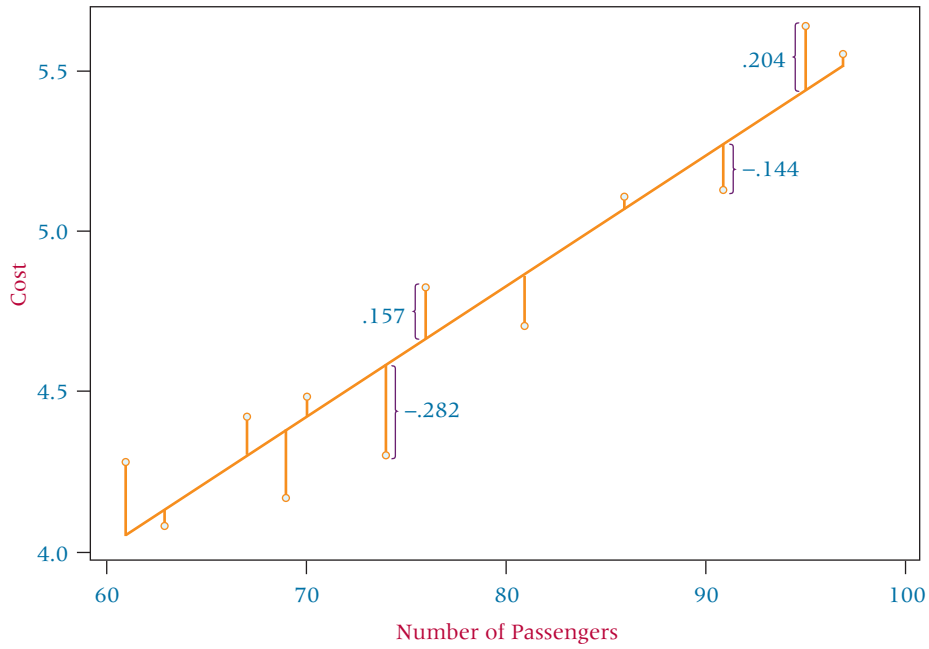
Note that the sum of the residuals is approximately zero. Except for rounding error, the sum of the residuals is *always zero*. The reason is that a residual is geometrically the vertical distance from the regression line to a data point. The equations used to solve for the slope

TABLE 12.5
Predicted Values and Residuals for the Airline Cost Example

Number of Passengers x	Cost ($1,000) y	Predicted Value $\hat{y}$	Residual $y - \hat{y}$
61	4.280	4.053	.227
63	4.080	4.134	−.054
67	4.420	4.297	.123
69	4.170	4.378	−.208
70	4.480	4.419	.061
74	4.300	4.582	−.282
76	4.820	4.663	.157
81	4.700	4.867	−.167
86	5.110	5.070	.040
91	5.130	5.274	−.144
95	5.640	5.436	.204
97	5.560	5.518	.042
		$\Sigma(y - \hat{y}) =$	−.001

FIGURE 12.7

Close-Up Minitab Scatter Plot with Residuals for the Airline Cost Example



and intercept place the line geometrically in the middle of all points. Therefore, vertical distances from the line to the points will cancel each other and sum to zero. Figure 12.7 is a Minitab-produced scatter plot of the data and the residuals for the airline cost example.

An examination of the residuals may give the researcher an idea of how well the regression line fits the historical data points. The largest residual for the airline cost example is −.282, and the smallest is .040. Because the objective of the regression analysis was to predict the cost of flight in $1,000s, the regression line produces an error of $282 when there are 74 passengers and an error of only $40 when there are 86 passengers. This result presents the *best* and *worst* cases for the residuals. The researcher must examine other residuals to determine how well the regression model fits other data points.

Sometimes residuals are used to locate outliers. **Outliers** are *data points that lie apart from the rest of the points.* Outliers can produce residuals with large magnitudes and are usually easy to identify on scatter plots. Outliers can be the result of misrecorded or miscoded data, or they may simply be data points that do not conform to the general trend. The equation of the regression line is influenced by every data point used in its calculation in a manner similar to the arithmetic mean. Therefore, outliers sometimes can unduly influence the regression line by "pulling" the line toward the outliers. The origin of outliers must be investigated to determine whether they should be retained or whether the regression equation should be recomputed without them.

Residuals are usually plotted against the *x*-axis, which reveals a view of the residuals as *x* increases. Figure 12.8 shows the residuals plotted by Excel against the *x*-axis for the airline cost example.

FIGURE 12.8

Excel Graph of Residuals for the Airline Cost Example

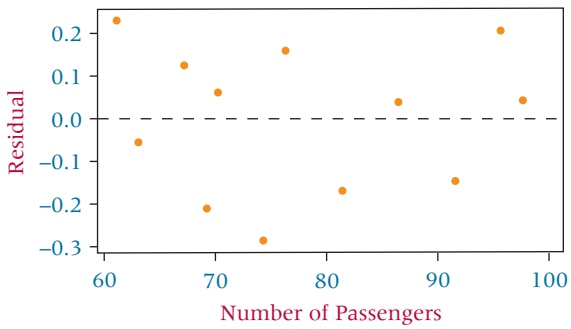


FIGURE 12.9
Nonlinear Residual Plot

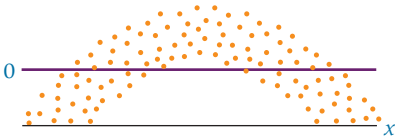
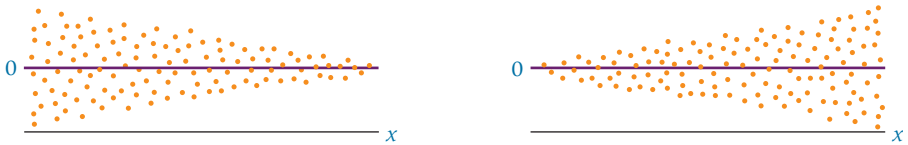


FIGURE 12.10
Nonconstant Error Variance



Using Residuals to Test the Assumptions of the Regression Model

One of the major uses of residual analysis is to test some of the assumptions underlying regression. The following are the assumptions of simple regression analysis.

1. The model is linear.
2. The error terms have constant variances.
3. The error terms are independent.
4. The error terms are normally distributed.

A particular method for studying the behavior of residuals is the residual plot. The **residual plot** is *a type of graph in which the residuals for a particular regression model are plotted along with their associated value of x as an ordered pair* $(x, y - \hat{y})$. Information about how well the regression assumptions are met by the particular regression model can be gleaned by examining the plots. Residual plots are more meaningful with larger sample sizes. For small sample sizes, residual plot analyses can be problematic and subject to over-interpretation. Hence, because the airline cost example is constructed from only 12 pairs of data, one should be cautious in reaching conclusions from Figure 12.8. The residual plots in Figures 12.9, 12.10, and 12.11, however, represent large numbers of data points and therefore are more likely to depict overall trends accurately.

If a residual plot such as the one in Figure 12.9 appears, the assumption that the model is linear does not hold. Note that the residuals are negative for low and high values of x and are positive for middle values of x. The graph of these residuals is parabolic, not linear. The residual plot does not have to be shaped in this manner for a nonlinear relationship to exist. Any significant deviation from an approximately linear residual plot may mean that a nonlinear relationship exists between the two variables.

The assumption of *constant error variance* sometimes is called **homoscedasticity**. If *the error variances are not constant* (called **heteroscedasticity**), the residual plots might look like one of the two plots in Figure 12.10. Note in Figure 12.10(a) that the error variance is greater for small values of x and smaller for large values of x. The situation is reversed in Figure 12.10(b).

If the error terms are not independent, the residual plots could look like one of the graphs in Figure 12.11. According to these graphs, instead of each error term being independent of the one next to it, the value of the residual is a function of the residual value next to it. For example, a large positive residual is next to a large positive residual and a small negative residual is next to a small negative residual.

The graph of the residuals from a regression analysis that meets the assumptions—a *healthy residual graph*—might look like the graph in Figure 12.12. The plot is relatively linear; the variances of the errors are about equal for each value of x, and the error terms do not appear to be related to adjacent terms.

FIGURE 12.11
Graphs of Nonindependent Error Terms

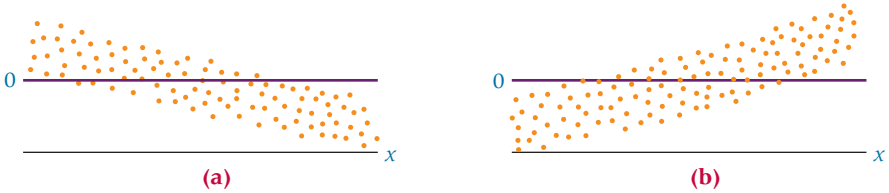
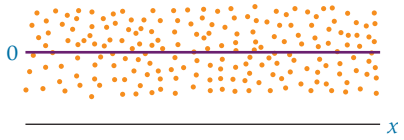


FIGURE 12.12
Healthy Residual Graph

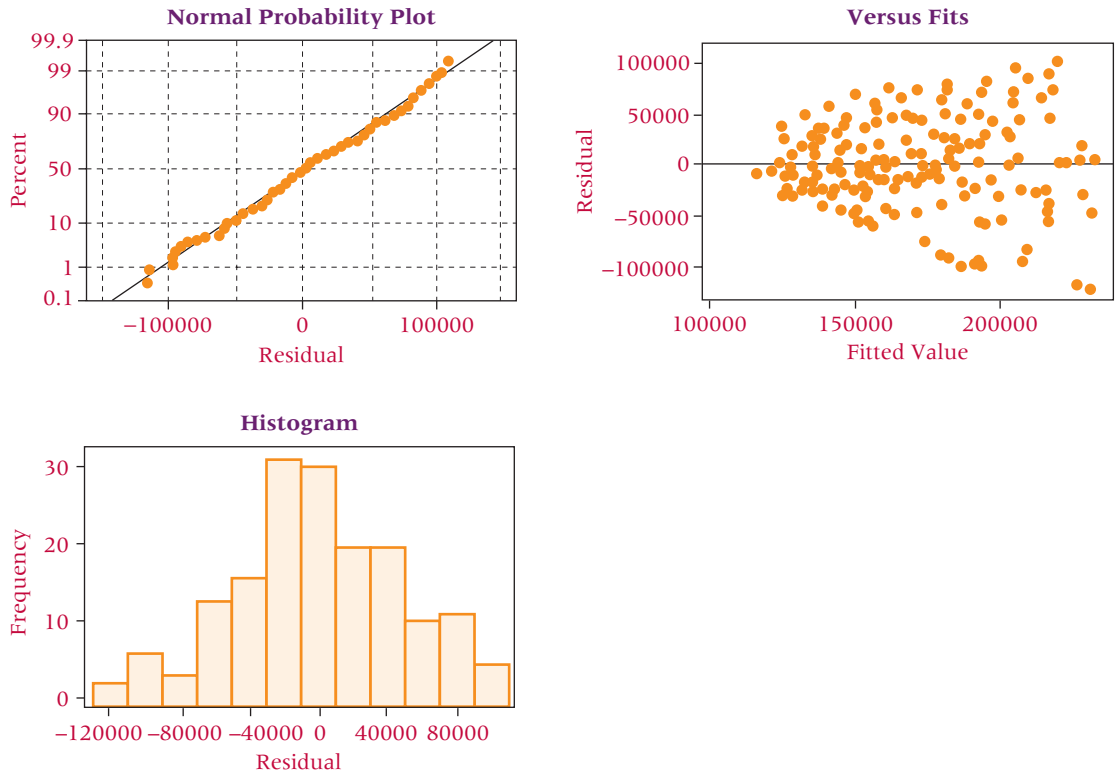


Using the Computer for Residual Analysis

Some computer programs contain mechanisms for analyzing residuals for violations of the regression assumptions. Minitab has the capability of providing graphical analysis of residuals. Figure 12.13 displays Minitab's residual graphic analyses for a regression model developed to predict the production of carrots in the United States per month by the total production of sweet corn. The data were gathered over a time period of 168 consecutive months (see WileyPLUS for the agricultural database).

These Minitab residual model diagnostics consist of three different plots. The graph on the upper right is a plot of the residuals versus the fits. Note that this residual plot "flares-out" as x gets larger. This pattern is an indication of heteroscedasticity, which is a violation of the assumption of constant variance for error terms. The graph in the upper left is a normal probability plot of the residuals. A straight line indicates that the residuals are normally distributed. Observe that this normal plot is relatively close to being a straight line, indicating that the residuals are nearly normal in shape. This normal distribution is confirmed by the graph on the lower left, which is a histogram of the residuals. The histogram groups residuals in classes so the researcher can observe where groups of the residuals lie without having to rely on the residual plot and to validate the notion that the residuals are approximately normally distributed. In this problem, the pattern is indicative of at least a mound-shaped distribution of residuals.

FIGURE 12.13
Minitab Residual Analyses



DEMONSTRATION PROBLEM 12.2

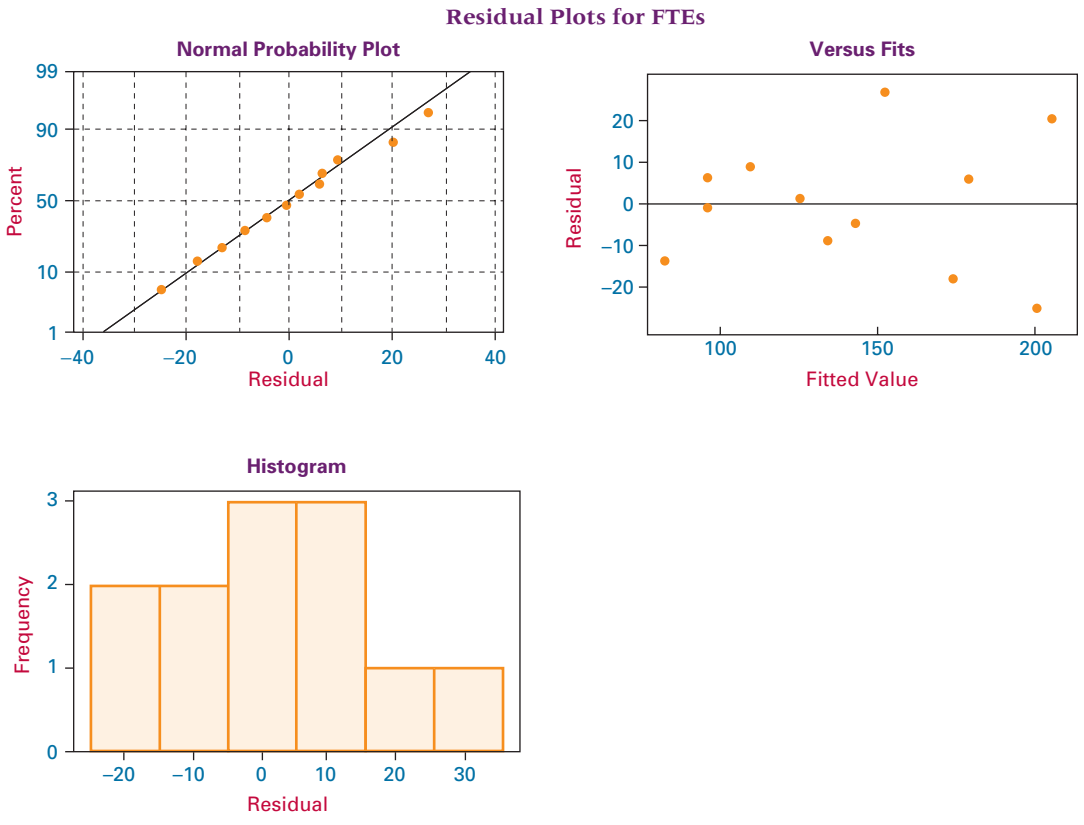
Compute the residuals for Demonstration Problem 12.1 in which a regression model was developed to predict the number of full-time equivalent workers (FTEs) by the number of beds in a hospital. Analyze the residuals by using Minitab graphic diagnostics.

Solution

The data and computed residuals are shown in the following table.

Hospital	Number of Beds x	FTES y	Predicted Value $\hat{y}$	Residuals $y - \hat{y}$
1	23	69	82.22	−13.22
2	29	95	95.62	−.62
3	29	102	95.62	6.38
4	35	118	109.01	8.99
5	42	126	124.63	1.37
6	46	125	133.56	−8.56
7	50	138	142.49	−4.49
8	54	178	151.42	26.58
9	64	156	173.74	−17.74
10	66	184	178.20	5.80
11	76	176	200.52	−24.52
12	78	225	204.98	20.02
				$\Sigma(y - \hat{y}) = -.01$

Note that the regression model fits these particular data well for hospitals 2 and 5, as indicated by residuals of −.62 and 1.37 FTEs, respectively. For hospitals 1, 8, 9, 11, and 12, the residuals are relatively large, indicating that the regression model does



not fit the data for these hospitals well. The Residuals Versus the Fitted Values graph indicates that the residuals seem to increase as x increases, indicating a potential problem with heteroscedasticity. The normal plot of residuals indicates that the residuals are nearly normally distributed. The histogram of residuals shows that the residuals pile up in the middle, but are somewhat skewed toward the larger positive values.

12.4 PROBLEMS

12.13 Determine the equation of the regression line for the following data, and compute the residuals.

x	15	8	19	12	5
y	47	36	56	44	21

12.14 Solve for the predicted values of y and the residuals for the data in Problem 12.6. The data are provided here again:

x	12	21	28	8	20
y	17	15	22	19	24

12.15 Solve for the predicted values of y and the residuals for the data in Problem 12.7. The data are provided here again:

x	140	119	103	91	65	29	24
y	25	29	46	70	88	112	128

12.16 Solve for the predicted values of y and the residuals for the data in Problem 12.8. The data are provided here again:

Advertising	12.5	3.7	21.6	60.0	37.6	6.1	16.8	41.2
Sales	148	55	338	994	541	89	126	379

12.17 Solve for the predicted values of y and the residuals for the data in Problem 12.9. The data are provided here again:

Bond Rate	5%	12%	9%	15%	7%
Prime Interest Rate	16%	6%	8%	4%	7%

12.18 In problem 12.10, you were asked to develop the equation of a regression model to predict the number of business bankruptcies by the number of firm births. Using this regression model and the data given in problem 12.10 (and provided here again), solve for the predicted values of y and the residuals. Comment on the size of the residuals.

Business Bankruptcies (1,000)	Firm Births (10,000)
34.3	58.1
35.0	55.4
38.5	57.0
40.1	58.5
35.5	57.4
37.9	58.0

12.19 The equation of a regression line is

$$\hat{y} = 50.506 - 1.646x$$

and the data are as follows.

x	5	7	11	12	19	25
y	47	38	32	24	22	10

Solve for the residuals and graph a residual plot. Do these data seem to violate any of the assumptions of regression?

12.20 Wisconsin is an important milk-producing state. Some people might argue that because of transportation costs, the cost of milk increases with the distance of markets from Wisconsin. Suppose the milk prices in eight cities are as follows.

Cost of Milk (per gallon)	Distance from Madison (miles)
$2.64	1,245
2.31	425
2.45	1,346
2.52	973
2.19	255
2.55	865
2.40	1,080
2.37	296

Use the prices along with the distance of each city from Madison, Wisconsin, to develop a regression line to predict the price of a gallon of milk by the number of miles the city is from Madison. Use the data and the regression equation to compute residuals for this model. Sketch a graph of the residuals in the order of the x values. Comment on the shape of the residual graph.

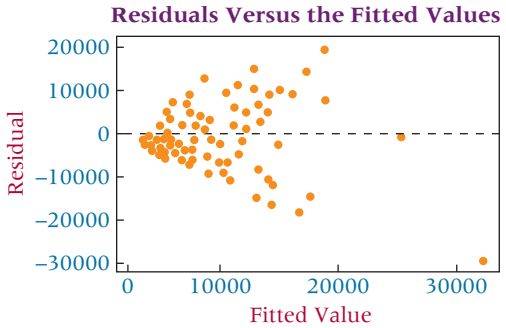
12.21 Graph the following residuals, and indicate which of the assumptions underlying regression appear to be in jeopardy on the basis of the graph.

x	$y - \hat{y}$
213	−11
216	−5
227	−2
229	−1
237	+6
247	+10
263	+12

12.22 Graph the following residuals, and indicate which of the assumptions underlying regression appear to be in jeopardy on the basis of the graph.

x	$y - \hat{y}$
10	+6
11	+3
12	−1
13	−11
14	−3
15	+2
16	+5
17	+8

12.23 Study the following Minitab Residuals Versus Fits graphic for a simple regression analysis. Comment on the residual evidence of lack of compliance with the regression assumptions.



12.5 STANDARD ERROR OF THE ESTIMATE

Residuals represent errors of estimation for individual points. With large samples of data, residual computations become laborious. Even with computers, a researcher sometimes has difficulty working through pages of residuals in an effort to understand the error of the regression model. An alternative way of examining the error of the model is the standard error of the estimate, which provides a single measurement of the regression error.

Because the sum of the residuals is zero, attempting to determine the total amount of error by summing the residuals is fruitless. This zero-sum characteristic of residuals can be avoided by squaring the residuals and then summing them.

Table 12.6 contains the airline cost data from Table 12.3, along with the residuals and the residuals squared. The *total of the residuals squared* column is called the **sum of squares of error (SSE)**.

SUM OF SQUARES OF ERROR

$$\text{SSE} = \Sigma(y - \hat{y})^2$$

In theory, infinitely many lines can be fit to a sample of points. However, formulas 12.2 and 12.4 produce a line of best fit for which the SSE is the smallest for any line that can be fit to the sample data. This result is guaranteed, because formulas 12.2 and 12.4 are derived from calculus to minimize SSE. For this reason, the regression process used in this chapter is called *least squares* regression.

A computational version of the equation for computing SSE is less meaningful in terms of interpretation than $\Sigma(y - \hat{y})^2$ but it is usually easier to compute. The computational formula for SSE follows.

COMPUTATIONAL FORMULA FOR SSE

$$\text{SSE} = \Sigma y^2 - b_0\Sigma y - b_1\Sigma xy$$

For the airline cost example,

$$\Sigma y^2 = \Sigma[(4.280)^2 + (4.080)^2 + (4.420)^2 + (4.170)^2 + (4.480)^2 + (4.300)^2 + (4.820)^2 + (4.700)^2 + (5.110)^2 + (5.130)^2 + (5.640)^2 + (5.560)^2] = 270.9251$$

$$b_0 = 1.5697928$$

TABLE 12.6
Determining SSE for the Airline Cost Example

Number of Passengers x	Cost ($1,000) y	Residual $y - \hat{y}$	$(y - \hat{y})^2$
61	4.280	.227	.05153
63	4.080	−.054	.00292
67	4.420	.123	.01513
69	4.170	−.208	.04326
70	4.480	.061	.00372
74	4.300	−.282	.07952
76	4.820	.157	.02465
81	4.700	−.167	.02789
86	5.110	.040	.00160
91	5.130	−.144	.02074
95	5.640	.204	.04162
97	5.560	.042	.00176
		$\Sigma(y - \hat{y}) = -.001$	$\Sigma(y - \hat{y})^2 = .31434$

Sum of squares of error = SSE = .31434

$$b_1 = .0407016^*$$
$$\Sigma y = 56.69$$
$$\Sigma xy = 4462.22$$
$$\text{SSE} = \Sigma y^2 - b_0\Sigma y - b_1\Sigma xy$$
$$= 270.9251 - (1.5697928)(56.69) - (.0407016)(4462.22) = .31405$$

The slight discrepancy between this value and the value computed in Table 12.6 is due to rounding error.

The sum of squares error is in part a function of the number of pairs of data being used to compute the sum, which lessens the value of SSE as a measurement of error. A more useful measurement of error is the standard error of the estimate. The **standard error of the estimate**, denoted s_e, is *a standard deviation of the error of the regression model* and has a more practical use than SSE. The standard error of the estimate follows.

STANDARD ERROR OF THE ESTIMATE

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}}$$

The standard error of the estimate for the airline cost example is

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{.31434}{10}} = .1773$$

How is the standard error of the estimate used? As previously mentioned, the standard error of the estimate is a standard deviation of error. Recall from Chapter 3 that if data are approximately normally distributed, the empirical rule states that about 68% of all values are within $\mu \pm 1\sigma$ and that about 95% of all values are within $\mu \pm 2\sigma$. One of the assumptions for regression states that for a given x the error terms are normally distributed. Because the error terms are normally distributed, s_e is the standard deviation of error, and the average error is zero, approximately 68% of the error values (residuals) should be within $0 \pm 1s_e$ and 95% of the error values (residuals) should be within $0 \pm 2s_e$. By having knowledge of the variables being studied and by examining the value of s_e, the researcher can often make a judgment about the fit of the regression model to the data by using s_e. How can the s_e value for the airline cost example be interpreted?

The regression model in that example is used to predict airline cost by number of passengers. Note that the range of the airline cost data in Table 12.3 is from 4.08 to 5.64 ($4,080 to $5,640). The regression model for the data yields an s_e of .1773. An interpretation of s_e is that the standard deviation of error for the airline cost example is $177.30. If the error terms were normally distributed about the given values of x, approximately 68% of the error terms would be within ±$177.30 and 95% would be within ±2($177.30) = ±$354.60. Examination of the residuals reveals that 100% of the residuals are within $2s_e$. The standard error of the estimate provides a single measure of error, which, if the researcher has enough background in the area being analyzed, can be used to understand the magnitude of errors in the model. In addition, some researchers use the standard error of the estimate to identify outliers. They do so by looking for data that are outside $\pm 2s_e$ or $\pm 3s_e$.

DEMONSTRATION PROBLEM 12.3

Compute the sum of squares of error and the standard error of the estimate for Demonstration Problem 12.1, in which a regression model was developed to predict the number of FTEs at a hospital by the number of beds.

*__Note:__ In previous sections, the values of the slope and intercept were rounded off for ease of computation and interpretation. They are shown here with more precision in an effort to reduce rounding error.

Solution

Hospital	Number of Beds x	FTES y	Residuals $y - \hat{y}$	$(y - \hat{y})^2$
1	23	69	−13.22	174.77
2	29	95	−.62	−0.38
3	29	102	6.38	40.70
4	35	118	8.99	80.82
5	42	126	1.37	1.88
6	46	125	−8.56	73.27
7	50	138	−4.49	20.16
8	54	178	26.58	706.50
9	64	156	−17.74	314.71
10	66	184	5.80	33.64
11	76	176	−24.52	601.23
12	78	225	20.02	400.80
	$\Sigma x = 592$	$\Sigma y = 1692$	$\Sigma(y - \hat{y}) = -.01$	$\Sigma(y - \hat{y})^2 = 2448.86$

SSE = 2448.86

$$S_e = \sqrt{\frac{\text{SSE}}{n - 2}} = \sqrt{\frac{2448.86}{10}} = 15.65$$

The standard error of the estimate is 15.65 FTEs. An examination of the residuals for this problem reveals that 8 of 12 (67%) are within $\pm 1 s_e$ and 100% are within $\pm 2 s_e$. Is this size of error acceptable? Hospital administrators probably can best answer that question.

12.5 PROBLEMS

12.24 Determine the sum of squares of error (SSE) and the standard error of the estimate (s_e) for Problem 12.6. Determine how many of the residuals computed in Problem 12.14 (for Problem 12.6) are within one standard error of the estimate. If the error terms are normally distributed, approximately how many of these residuals should be within $\pm 1 s_e$?

12.25 Determine the SSE and the s_e for Problem 12.7. Use the residuals computed in Problem 12.15 (for Problem 12.7) and determine how many of them are within $\pm 1 s_e$ and $\pm 2 s_e$. How do these numbers compare with what the empirical rule says should occur if the error terms are normally distributed?

12.26 Determine the SSE and the s_e for Problem 12.8. Think about the variables being analyzed by regression in this problem and comment on the value of s_e.

12.27 Determine the SSE and s_e for Problem 12.9. Examine the variables being analyzed by regression in this problem and comment on the value of s_e.

12.28 In problem 12.10, you were asked to develop the equation of a regression model to predict the number of business bankruptcies by the number of firm births. For this regression model, solve for the standard error of the estimate and comment on it.

12.29 Use the data from problem 12.19 and determine the s_e.

12.30 Determine the SSE and the s_e for Problem 12.20. Comment on the size of s_e for this regression model, which is used to predict the cost of milk.

12.31 Determine the equation of the regression line to predict annual sales of a company from the yearly stock market volume of shares sold in a recent year. Compute the standard error of the estimate for this model. Does volume of shares sold appear to be a good predictor of a company's sales? Why or why not?

Company	Annual Sales ($ billions)	Annual Volume (millions of shares)
Merck	10.5	728.6
Altria	48.1	497.9
IBM	64.8	439.1
Eastman Kodak	20.1	377.9
Bristol-Myers Squibb	11.4	375.5
General Motors	123.8	363.8
Ford Motors	89.0	276.3

12.6 COEFFICIENT OF DETERMINATION

A widely used measure of fit for regression models is the **coefficient of determination**, or r^2. The coefficient of determination is *the proportion of variability of the dependent variable (y) accounted for or explained by the independent variable (x).*

The coefficient of determination ranges from 0 to 1. An r^2 of zero means that the predictor accounts for none of the variability of the dependent variable and that there is no regression prediction of y by x. An r^2 of 1 means perfect prediction of y by x and that 100% of the variability of y is accounted for by x. Of course, most r^2 values are between the extremes. The researcher must interpret whether a particular r^2 is high or low, depending on the use of the model and the context within which the model was developed.

In exploratory research where the variables are less understood, low values of r^2 are likely to be more acceptable than they are in areas of research where the parameters are more developed and understood. One NASA researcher who uses vehicular weight to predict mission cost searches for the regression models to have an r^2 of .90 or higher. However, a business researcher who is trying to develop a model to predict the motivation level of employees might be pleased to get an r^2 near .50 in the initial research.

The dependent variable, y, being predicted in a regression model has a variation that is measured by the sum of squares of y (SS_{yy}):

$$SS_{yy} = \Sigma(y - \bar{y})^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

and is the sum of the squared deviations of the y values from the mean value of y. This variation can be broken into two additive variations: the *explained variation,* measured by the sum of squares of regression (SSR), and the *unexplained variation,* measured by the sum of squares of error (SSE). This relationship can be expressed in equation form as

$$SS_{yy} = SSR + SSE$$

If each term in the equation is divided by SS_{yy}, the resulting equation is

$$1 = \frac{SSR}{SS_{yy}} + \frac{SSE}{SS_{yy}}$$

The term r^2 is the proportion of the y variability that is explained by the regression model and represented here as

$$r^2 = \frac{SSR}{SS_{yy}}$$

Substituting this equation into the preceding relationship gives

$$1 = r^2 + \frac{SSE}{SS_{yy}}$$

Solving for r^2 yields formula 12.5.

COEFFICIENT OF DETERMINATION (12.5)

$$r^2 = 1 - \frac{\text{SSE}}{\text{SS}_{yy}} = 1 - \frac{\text{SSE}}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}$$

Note: $0 \leq r^2 \leq 1$

The value of r^2 for the airline cost example is solved as follows:

$$\text{SSE} = .31434$$

$$\text{SS}_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 270.9251 - \frac{(56.69)^2}{12} = 3.11209$$

$$r^2 = 1 - \frac{\text{SSE}}{\text{SS}_{yy}} = 1 - \frac{.31434}{3.11209} = .899$$

That is, 89.9% of the variability of the cost of flying a Boeing 737 airplane on a commercial flight is explained by variations in the number of passengers. This result also means that 11.1% of the variance in airline flight cost, y, is unaccounted for by x or unexplained by the regression model.

The coefficient of determination can be solved for directly by using

$$r^2 = \frac{\text{SSR}}{\text{SS}_{yy}}$$

It can be shown through algebra that

$$\text{SSR} = b_1^2 \text{SS}_{xx}$$

From this equation, a computational formula for r^2 can be developed.

COMPUTATIONAL FORMULA FOR r^2

$$r^2 = \frac{b_1^2 SS_{xx}}{\text{SS}_{yy}}$$

For the airline cost example, $b_1 = .0407016$, $\text{SS}_{xx} = 1689$, and $\text{SS}_{yy} = 3.11209$. Using the computational formula for r^2 yields

$$r^2 = \frac{(.0407016)^2(1689)}{3.11209} = .899$$

DEMONSTRATION PROBLEM 12.4

Compute the coefficient of determination (r^2) for Demonstration Problem 12.1, in which a regression model was developed to predict the number of FTEs of a hospital by the number of beds.

Solution

$$\text{SSE} = 2448.86$$

$$\text{SS}_{yy} = 260{,}136 - \frac{(1692)^2}{12} = 21{,}564$$

$$r^2 = 1 - \frac{\text{SSE}}{\text{SS}_{yy}} = 1 - \frac{2448.86}{21{,}564} = .886$$

This regression model accounts for 88.6% of the variance in FTEs, leaving only 11.4% unexplained variance.

Using $\text{SS}_{xx} = 3838.667$ and $b_1 = 2.232$ from Demonstration Problem 12.1, we can solve for r^2 with the computational formula:

$$r^2 = \frac{b_1^2 \text{SS}_{xx}}{\text{SS}_{yy}} = \frac{(2.232)^2(3838.667)}{21{,}564} = .886$$

Relationship Between r and r^2

Is r, the coefficient of correlation (introduced in Section 12.1), related to r^2, the coefficient of determination in linear regression? The answer is yes: r^2 equals $(r)^2$. The coefficient of determination is the square of the coefficient of correlation. In Demonstration Problem 12.1, a regression model was developed to predict FTEs by number of hospital beds. The r^2 value for the model was .886. Taking the square root of this value yields $r = .941$, which is the correlation between the sample number of beds and FTEs. A word of caution here: Because r^2 is always positive, solving for r by taking $\sqrt{r^2}$ gives the correct magnitude of r but may give the wrong sign. The researcher must examine the sign of the slope of the regression line to determine whether a positive or negative relationship exists between the variables and then assign the appropriate sign to the correlation value.

12.6 PROBLEMS

12.32 Compute r^2 for Problem 12.24 (Problem 12.6). Discuss the value of r^2 obtained.

12.33 Compute r^2 for Problem 12.25 (Problem 12.7). Discuss the value of r^2 obtained.

12.34 Compute r^2 for Problem 12.26 (Problem 12.8). Discuss the value of r^2 obtained.

12.35 Compute r^2 for Problem 12.27 (Problem 12.9). Discuss the value of r^2 obtained.

12.36 In problem 12.10, you were asked to develop the equation of a regression model to predict the number of business bankruptcies by the number of firm births. For this regression model, solve for the coefficient of determination and comment on it.

12.37 The Conference Board produces a Consumer Confidence Index (CCI) that reflects people's feelings about general business conditions, employment opportunities, and their own income prospects. Some researchers may feel that consumer confidence is a function of the median household income. Shown here are the CCIs for nine years and the median household incomes for the same nine years published by the U.S. Census Bureau. Determine the equation of the regression line to predict the CCI from the median household income. Compute the standard error of the estimate for this model. Compute the value of r^2. Does median household income appear to be a good predictor of the CCI? Why or why not?

CCI	Median Household Income ($1,000)
116.8	37.415
91.5	36.770
68.5	35.501
61.6	35.047
65.9	34.700
90.6	34.942
100.0	35.887
104.6	36.306
125.4	37.005

12.7 HYPOTHESIS TESTS FOR THE SLOPE OF THE REGRESSION MODEL AND TESTING THE OVERALL MODEL

Testing the Slope

A hypothesis test can be conducted on the sample slope of the regression model to determine whether the population slope is significantly different from zero. This test is another way to determine how well a regression model fits the data. Suppose a researcher decides

that it is not worth the effort to develop a linear regression model to predict y from x. An alternative approach might be to average the y values and use $\bar{y}$ as the predictor of y for all values of x. For the airline cost example, instead of using number of passengers as the predictor, the researcher would use the average value of airline cost, $\bar{y}$, as the predictor. In this case the average value of y is

$$\bar{y} = \frac{56.69}{12} = 4.7242, \text{ or } \$4{,}724.20$$

Using this result as a model to predict y, if the number of passengers is 61, 70, or 95—or any other number—the predicted value of y is still 4.7242. Essentially, this approach fits the line of $\bar{y} = 4.7242$ through the data, which is a horizontal line with a slope of zero. Would a regression analysis offer anything more than the $\bar{y}$ model? Using this nonregression model (the $\bar{y}$ model) as a worst case, the researcher can analyze the regression line to determine whether it adds a more significant amount of predictability of y than does the $\bar{y}$ model. Because the slope of the $\bar{y}$ line is zero, one way to determine whether the regression line adds significant predictability is to test the population slope of the regression line to find out whether the slope is different from zero. As the slope of the regression line diverges from zero, the regression model is adding predictability that the $\bar{y}$ line is not generating. For this reason, testing the slope of the regression line to determine whether the slope is different from zero is important. If the slope is not different from zero, the regression line is doing nothing more than the $\bar{y}$ line in predicting y.

How does the researcher go about testing the slope of the regression line? Why not just examine the observed sample slope? For example, the slope of the regression line for the airline cost data is .0407. This value is obviously not zero. The problem is that this slope is obtained from a sample of 12 data points; and if another sample was taken, it is likely that a different slope would be obtained. For this reason, the population slope is statistically tested using the sample slope. The question is: If all the pairs of data points for the population were available, would the slope of that regression line be different from zero? Here the sample slope, b_1, is used as evidence to test whether the population slope is different from zero. The hypotheses for this test follow.

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

Note that this test is two tailed. The null hypothesis can be rejected if the slope is either negative or positive. A negative slope indicates an inverse relationship between x and y. That is, larger values of x are related to smaller values of y, and vice versa. Both negative and positive slopes can be different from zero. To determine whether there is a significant *positive* relationship between two variables, the hypotheses would be one tailed, or

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 > 0$$

To test for a significant *negative* relationship between two variables, the hypotheses also would be one tailed, or

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 < 0$$

In each case, testing the null hypothesis involves a t test of the slope.

t TEST OF SLOPE

$$t = \frac{b_1 - \beta_1}{s_b}$$

where

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}}$$

$$s_e = \sqrt{\frac{SSE}{n-2}}$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

β_1 = the hypothesized slope
df = $n - 2$

The test of the slope of the regression line for the airline cost regression model for $\alpha = .05$ follows. The regression line derived for the data is

$$\hat{y} = 1.57 + .0407x$$

The sample slope is $.0407 = b_1$. The value of s_e is .1773, $\Sigma x = 930$, $\Sigma x^2 = 73{,}764$, and $n = 12$. The hypotheses are

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

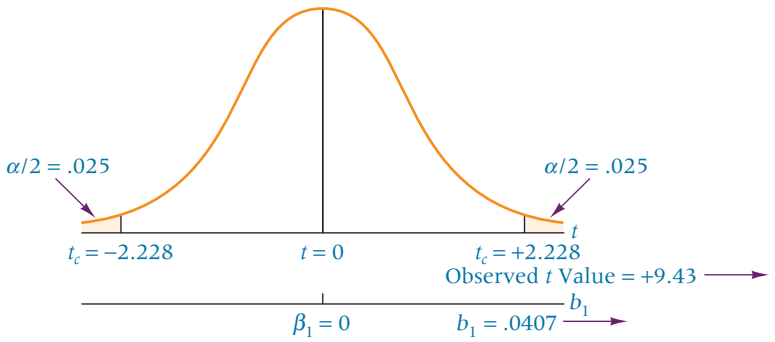
The df $= n - 2 = 12 - 2 = 10$. As this test is two tailed, $\alpha/2 = .025$. The table t value is $t_{.025,10} = \pm 2.228$. The observed t value for this sample slope is

$$t = \frac{.0407 - 0}{.1773\Big/\sqrt{73{,}764 - \frac{(930)^2}{12}}} = 9.43$$

As shown in Figure 12.14, the t value calculated from the sample slope falls in the rejection region and the p-value is .00000014. The null hypothesis that the population slope is zero is rejected. This linear regression model is adding significantly more predictive information to the $\bar{y}$ model (no regression).

It is desirable to reject the null hypothesis in testing the slope of the regression model. In rejecting the null hypothesis of a zero population slope, we are stating that the regression model is adding something to the explanation of the variation of the dependent variable that the average value of y model does not. Failure to reject the null hypothesis in this test causes the researcher to conclude that the regression model has no predictability of the dependent variable, and the model, therefore, has little or no use.

FIGURE 12.14
t Test of Slope from Airline Cost Example



STATISTICS IN BUSINESS TODAY

Predicting the Price of an SUV

What variables are good predictors of the base price of a new car? In a *Wall Street Journal* article on the Ford Expedition, data are displayed for five variables on five different makes of large SUVs. The variables are base price, engine horsepower, weight (in pounds), towing capacity (in pounds), and city EPA mileage. The SUV makes are Ford Expedition Eddie Bauer 4 × 4, Toyota Sequoia Limited, Chevrolet Tahoe LT, Acura MDX, and Dodge Durango R/T. The base prices of these five models ranged from $34,700 to $42,725. Suppose a business researcher wanted to develop a regression model to predict the base price of these cars. What variable would be the strongest predictor and how strong would the prediction be?

Using a correlation matrix constructed from the data for the five variables, it was determined that weight was most correlated with base price and had the greatest potential as a predictor. Towing capacity had the second highest correlation with base price, followed by city EPA mileage and horsepower. City EPA mileage was negatively related to base price, indicating that the more expensive SUVs tended to be "gas guzzlers."

A regression model was developed using weight as a predictor of base price. The Minitab output from the data follows. Excel output contains similar items.

Regression Analysis: Base Price Versus Weight

```
The regression equation is
Base Price = 10140 + 5.77 Weight
Predictor    Coef    SE Coef       T      P
Constant    10140       8473    1.20  0.317
Weight      5.769      1.679    3.44  0.041
S = 1699   R-Sq = 79.7%   R-Sq(adj) = 73.0%
```

Note that the r^2 for this model is almost 80% and that the t statistic is significant at $\alpha = .05$. In the regression equation, the slope indicates that for every pound of weight increase there is a $5.77 increase in the price. The y-intercept indicates that if the SUV weighed nothing at all, it would still cost $10,140! The standard error of the estimate is $1,699.

Regression models were developed for each of the other possible predictor variables. Towing capacity was the next best predictor variable producing an r^2 of 31.4%. City EPA mileage produced an r^2 of 20%, and horsepower produced an r^2 of 6.9%.

Source: Adapted from Jonathan Welsh, "The Biggest Station Wagon of Them All," *The Wall Street Journal*, June 7, 2002, p. W15C.

DEMONSTRATION PROBLEM 12.5

Test the slope of the regression model developed in Demonstration Problem 12.1 to predict the number of FTEs in a hospital from the number of beds to determine whether there is a significant positive slope. Use $\alpha = .01$.

Solution

The hypotheses for this problem are

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 > 0$$

The level of significance is .01. With 12 pairs of data, df = 10. The critical table t value is $t_{.01,10} = 2.764$. The regression line equation for this problem is

$$\hat{y} = 30.888 + 2.232x$$

The sample slope, b_1, is 2.232, and $s_e = 15.65$, $\Sigma x = 592$, $\Sigma x^2 = 33{,}044$, and $n = 12$. The observed t value for the sample slope is

$$t = \frac{2.232 - 0}{15.65 \Big/ \sqrt{33{,}044 - \dfrac{(592)^2}{12}}} = 8.84$$

The observed t value (8.84) is in the rejection region because it is greater than the critical table t value of 2.764 and the p-value is .0000024. The null hypothesis is rejected. The population slope for this regression line is significantly different from zero in the positive direction. This regression model is adding significant predictability over the $\bar{y}$ model.

Testing the Overall Model

It is common in regression analysis to compute an F test to determine the overall significance of the model. Most computer software packages include the F test and its associated ANOVA table as standard regression output. In multiple regression (Chapters 13 and 14), this test determines whether at least one of the regression coefficients (from multiple predictors) is different from zero. Simple regression provides only one predictor and only one regression coefficient to test. Because the regression coefficient is the slope of the regression line, the F test for overall significance is testing the same thing as the t test in simple regression. The hypotheses being tested in simple regression by the F test for overall significance are

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

In the case of simple regression analysis, $F = t^2$. Thus, for the airline cost example, the F value is

$$F = t^2 = (9.43)^2 = 88.92$$

The F value is computed directly by

$$F = \frac{SS_{reg}/df_{reg}}{SS_{err}/df_{err}} = \frac{MS_{reg}}{MS_{err}}$$

where

$df_{reg} = k$
$df_{err} = n - k - 1$
k = the number of independent variables

The values of the sum of squares (SS), degrees of freedom (df), and mean squares (MS) are obtained from the analysis of variance table, which is produced with other regression statistics as standard output from statistical software packages. Shown here is the analysis of variance table produced by Minitab for the airline cost example.

Analysis of Variance

Source	DF	SS	MS	F	p
Regression	1	2.7980	2.7980	89.09	0.000
Residual Error	10	0.3141	0.0314		
Total	11	3.1121			

The F value for the airline cost example is calculated from the analysis of variance table information as

$$F = \frac{2.7980/1}{.3141/10} = \frac{2.7980}{.03141} = 89.09$$

The difference between this value (89.09) and the value obtained by squaring the t statistic (88.92) is due to rounding error. The probability of obtaining an F value this large or larger by chance if there is no regression prediction in this model is .000 according to the ANOVA output (the p-value). This output value means it is highly unlikely that the population slope is zero and that there is no prediction due to regression from this model given the sample statistics obtained. Hence, it is highly likely that this regression model adds significant predictability of the dependent variable.

Note from the ANOVA table that the degrees of freedom due to regression are equal to 1. Simple regression models have only one independent variable; therefore, $k = 1$. The degrees of freedom error in simple regression analysis is always $n - k - 1 = n - 1 - 1 = n - 2$. With the degrees of freedom due to regression (1) as the numerator degrees of freedom and the degrees of freedom due to error $(n - 2)$ as the denominator degrees of freedom, Table A.7 can be used to obtain the critical F value ($F_{\alpha,1,n-2}$) to help make the hypothesis testing

decision about the overall regression model if the p-value of F is not given in the computer output. This critical F value is always found in the right tail of the distribution. In simple regression, the relationship between the critical t value to test the slope and the critical F value of overall significance is

$$t^2_{\alpha/2,n-2} = F_{\alpha,1,n-2}$$

For the airline cost example with a two-tailed test and $\alpha = .05$, the critical value of $t_{.025,10}$ is ± 2.228 and the critical value of $F_{.05,1,10}$ is 4.96.
$t^2_{.025,10} = (\pm 2.228)^2 = 4.96 = F_{.05,1,10}$

12.7 PROBLEMS

12.38 Test the slope of the regression line determined in Problem 12.6. Use $\alpha = .05$.

12.39 Test the slope of the regression line determined in Problem 12.7. Use $\alpha = .01$.

12.40 Test the slope of the regression line determined in Problem 12.8. Use $\alpha = .10$.

12.41 Test the slope of the regression line determined in Problem 12.9. Use a 5% level of significance.

12.42 Test the slope of the regression line developed in Problem 12.10. Use a 5% level of significance.

12.43 Study the following analysis of variance table, which was generated from a simple regression analysis. Discuss the F test of the overall model. Determine the value of t and test the slope of the regression line.

Analysis of Variance

Source	DF	SS	MS	F	p
Regression	1	116.65	116.65	8.26	0.021
Error	8	112.95	14.12		
Total	9	229.60			

12.8 ESTIMATION

One of the main uses of regression analysis is as a prediction tool. If the regression function is a good model, the researcher can use the regression equation to determine values of the dependent variable from various values of the independent variable. For example, financial brokers would like to have a model with which they could predict the selling price of a stock on a certain day by a variable such as unemployment rate or producer price index. Marketing managers would like to have a site location model with which they could predict the sales volume of a new location by variables such as population density or number of competitors. The airline cost example presents a regression model that has the potential to predict the cost of flying an airplane by the number of passengers.

In simple regression analysis, a point estimate prediction of y can be made by substituting the associated value of x into the regression equation and solving for y. From the airline cost example, if the number of passengers is $x = 73$, the predicted cost of the airline flight can be computed by substituting the x value into the regression equation determined in Section 12.3:

$$\hat{y} = 1.57 + .0407x = 1.57 + .0407(73) = 4.5411$$

The point estimate of the predicted cost is 4.5411 or $4,541.10.

Confidence Intervals to Estimate the Conditional Mean of y: $\mu_{y|x}$

Although a point estimate is often of interest to the researcher, the regression line is determined by a sample set of points; and if a different sample is taken, a different line will

result, yielding a different point estimate. Hence computing a *confidence interval* for the estimation is often useful. Because for any value of x (independent variable) there can be many values of y (dependent variable), one type of **confidence interval** is *an estimate of the average value of y for a given x.* This average value of y is denoted $E(y_x)$—the expected value of y and can be computed using formula (12.6).

CONFIDENCE INTERVAL TO ESTIMATE $E(y_x)$ FOR A GIVEN VALUE OF x (12.6)

$$\hat{y} \pm t_{\alpha/2,n-2} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

where

x_0 = a particular value of x

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

The application of this formula can be illustrated with construction of a 95% confidence interval to estimate the average value of y (airline cost) for the airline cost example when x (number of passengers) is 73. For a 95% confidence interval, $\alpha = .05$ and $\alpha/2 = .025$. The df $= n - 2 = 12 - 2 = 10$. The table t value is $t_{.025,10} = 2.228$. Other needed values for this problem, which were solved for previously, are

$$s_e = .1773 \quad \Sigma x = 930 \quad \bar{x} = 77.5 \quad \Sigma x^2 = 73{,}764$$

For $x_0 = 73$, the value of $\hat{y}$ is 4.5411. The computed confidence interval for the average value of y, $E(y_{73})$, is

$$4.5411 \pm (2.228)(.1773)\sqrt{\frac{1}{12} + \frac{(73 - 77.5)^2}{73{,}764 - \frac{(930)^2}{12}}} = 4.5411 \pm .1220$$

$$4.4191 \leq E(y_{73}) \leq 4.6631$$

That is, with 95% confidence the average value of y for $x = 73$ is between 4.4191 and 4.6631.

Table 12.7 shows confidence intervals computed for the airline cost example for several values of x to estimate the average value of y. Note that as x values get farther from the mean x value (77.5), the confidence intervals get wider; as the x values get closer to the mean, the confidence intervals narrow. The reason is that the numerator of the second term under the radical sign approaches zero as the value of x nears the mean and increases as x departs from the mean.

Prediction Intervals to Estimate a Single Value of y

A second type of interval in regression estimation is a **prediction interval** to *estimate a single value of y for a given value of x.*

TABLE 12.7

Confidence Intervals to Estimate the Average Value of y for Some x Values in the Airline Cost Example

x	Confidence Interval	
62	4.0934 ± .1876	3.9058 to 4.2810
68	4.3376 ± .1461	4.1915 to 4.4837
73	4.5411 ± .1220	4.4191 to 4.6631
85	5.0295 ± .1349	4.8946 to 5.1644
90	5.2230 ± .1656	5.0574 to 5.3886

PREDICTION INTERVAL TO ESTIMATE y FOR A GIVEN VALUE OF x (12.7)

$$\hat{y} \pm t_{\alpha/2,n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

where

x_0 = a particular value of x

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

Formula 12.7 is virtually the same as formula 12.6, except for the additional value of 1 under the radical. This additional value widens the prediction interval to estimate a single value of y from the confidence interval to estimate the average value of y. This result seems logical because the average value of y is toward the middle of a group of y values. Thus the confidence interval to estimate the average need not be as wide as the prediction interval produced by formula 12.7, which takes into account all the y values for a given x.

A 95% prediction interval can be computed to estimate the single value of y for $x = 73$ from the airline cost example by using formula 12.7. The same values used to construct the confidence interval to estimate the average value of y are used here.

$$t_{.025,10} = 2.228,\ s_e = .1773,\ \Sigma x = 930,\ \bar{x} = 77.5,\ \Sigma x^2 = 73{,}764$$

For $x_0 = 73$, the value of $\hat{y} = 4.5411$. The computed prediction interval for the single value of y is

$$4.5411 \pm (2.228)(.1773)\sqrt{1 + \frac{1}{12} + \frac{(73 - 77.5)^2}{73{,}764 - \frac{(930)^2}{12}}} = 4.5411 \pm .4134$$

$$4.1277 \leq y \leq 4.9545$$

Prediction intervals can be obtained by using the computer. Shown in Figure 12.15 is the computer output for the airline cost example. The output displays the predicted value for $x = 73$ ($\hat{y} = 4.5411$), a 95% confidence interval for the average value of y for $x = 73$, and a 95% prediction interval for a single value of y for $x = 73$. Note that the resulting values are virtually the same as those calculated in this section.

Figure 12.16 displays Minitab confidence intervals for various values of x for the average y value and the prediction intervals for a single y value for the airline example. Note that the intervals flare out toward the ends, as the values of x depart from the average x value. Note also that the intervals for a single y value are always wider than the intervals for the average y value for any given value of x.

An examination of the prediction interval formula to estimate y for a given value of x explains why the intervals flare out.

$$\hat{y} \pm t_{\alpha/2,n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

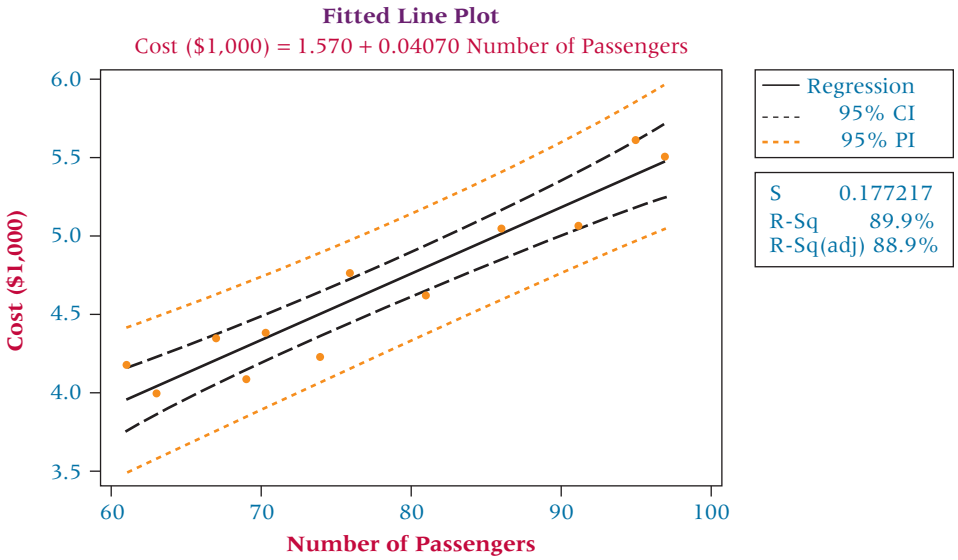
As we enter different values of x_0 from the regression analysis into the equation, the only thing that changes in the equation is $(x_0 - \bar{x})^2$. This expression increases as individual values of x_0 get farther from the mean, resulting in an increase in the width of the interval. The interval is narrower for values of x_0 nearer $\bar{x}$ and wider for values of x_0 further from $\bar{x}$. A comparison of formulas 12.6 and 12.7 reveals them to be identical except that formula 12.7—to compute a prediction interval to estimate y for a given value of x—contains a 1 under the radical sign. This distinction ensures that formula 12.7 will yield wider intervals than 12.6 for otherwise identical data.

FIGURE 12.15
Minitab Output for Prediction Intervals

```
Fit        StDev Fit      95.0% CI              95.0 PI
4.5410     0.0547         (4.4191, 4.6629)      (4.1278, 4.9543)
```

FIGURE 12.16

Minitab Intervals for Estimation



Caution: *A regression line is determined from a sample of points. The line, the r^2, the s_e, and the confidence intervals change for different sets of sample points. That is, the linear relationship developed for a set of points does not necessarily hold for values of x outside the domain of those used to establish the model. In the airline cost example, the domain of x values (number of passengers) varied from 61 to 97. The regression model developed from these points may not be valid for flights of say 40, 50, or 100 because the regression model was not constructed with x values of those magnitudes. However, decision makers sometimes extrapolate regression results to values of x beyond the domain of those used to develop the formulas (often in time-series sales forecasting). Understanding the limitations of this type of use of regression analysis is essential.*

DEMONSTRATION PROBLEM 12.6

Construct a 95% confidence interval to estimate the average value of y (FTEs) for Demonstration Problem 12.1 when $x = 40$ beds. Then construct a 95% prediction interval to estimate the single value of y for $x = 40$ beds.

Solution

For a 95% confidence interval, $\alpha = .05$ $n = 12$, and df $= 10$. The table t value is $t_{.025,10} = 2.228$; $s_e = 15.65$, $\sum x = 592$, $\bar{x} = 49.33$, and $\sum x^2 = 33{,}044$. For $x_0 = 40$, $\hat{y} = 120.17$. The computed confidence interval for the average value of y is

$$120.17 \pm (2.228)(15.65)\sqrt{\frac{1}{12} + \frac{(40 - 49.33)^2}{33{,}044 - \frac{(592)^2}{12}}} = 120.17 \pm 11.35$$

$$108.82 \leq E(y_{40}) \leq 131.52$$

With 95% confidence, the statement can be made that the average number of FTEs for a hospital with 40 beds is between 108.82 and 131.52.

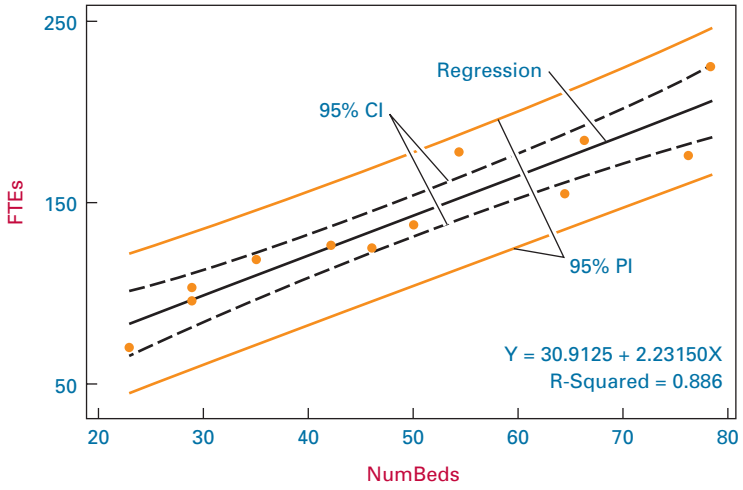
The computed prediction interval for the single value of y is

$$120.17 \pm (2.228)(15.65)\sqrt{1 + \frac{1}{12} + \frac{(40 - 49.33)^2}{33{,}044 - \frac{(592)^2}{12}}} = 120.17 \pm 36.67$$

$$83.5 \leq y \leq 156.84$$

With 95% confidence, the statement can be made that a single number of FTEs for a hospital with 40 beds is between 83.5 and 156.84. Obviously this interval is much wider than the 95% confidence interval for the average value of y for $x = 40$.

The following Minitab graph depicts the 95% interval bands for both the average y value and the single y values for all 12 x values in this problem. Note once again the flaring out of the bands near the extreme values of x.



12.8 PROBLEMS

12.44 Construct a 95% confidence interval for the average value of y for Problem 12.6. Use $x = 25$.

12.45 Construct a 90% prediction interval for a single value of y for Problem 12.7; use $x = 100$. Construct a 90% prediction interval for a single value of y for Problem 14.2; use $x = 130$. Compare the results. Which prediction interval is greater? Why?

12.46 Construct a 98% confidence interval for the average value of y for Problem 12.8; use $x = 20$. Construct a 98% prediction interval for a single value of y for Problem 14.3; use $x = 20$. Which is wider? Why?

12.47 Construct a 99% confidence interval for the average bond rate in Problem 12.9 for a prime interest rate of 10%. Discuss the meaning of this confidence interval.

12.9 USING REGRESSION TO DEVELOP A FORECASTING TREND LINE

Business researchers often use historical data with measures taken over time in an effort to forecast what might happen in the future. A particular type of data that often lends itself well to this analysis is **time-series data** defined as *data gathered on a particular characteristic over a period of time at regular intervals.* Some examples of time-series data are 10 years of weekly Dow Jones Industrial Averages, twelve months of daily oil production, or monthly consumption of coffee over a two-year period. To be useful to forecasters, time-series measurements need to be made in regular time intervals and arranged according to time of occurrence. As an example, consider the time-series sales data over a 10-year time period for the Huntsville Chemical Company shown in Table 12.8. Note that the measurements (sales) are taken over time and that the sales figures are given on a yearly basis. Time-series data can also be reported daily, weekly, monthly, quarterly, semi-annually, or for other defined time periods.

TABLE 12.8
Ten-Year Sales Data for Huntsville Chemicals

Year	Sales ($ millions)
2000	7.84
2001	12.26
2002	13.11
2003	15.78
2004	21.29
2005	25.68
2006	23.80
2007	26.43
2008	29.16
2009	33.06

It is generally believed that time-series data contain any one or combination of four elements: trend, cyclicality, seasonality, and irregularity. While each of these four elements will be discussed in greater deal in Chapter 15, Time-Series Forecasting and Index Numbers, here we examine **trend** and define it as *the long-term general direction of data.* Observing the scatter plot of the Huntsville Chemical Company's sales data shown in Figure 12.17, it is apparent that there is positive trend in the data. That is, there appears to be a long-term upward general direction of sales over time. How can trend be expressed in mathematical terms? In the field of forecasting, it is common to attempt to fit a trend line through time-series data by determining the equation of the trend line and then using the equation of the trend line to predict future data points. How does one go about developing such a line?

Determining the Equation of the Trend Line

Developing the equation of a linear trend line in forecasting is actually a special case of simple regression where the y or dependent variable is the variable of interest that a business analyst wants to forecast and for which a set of measurements has been taken over a period of time. For example, with the Huntsville Chemicals Company data, if company forecasters want to predict sales for the year 2012 using these data, sales would be the dependent variable in the simple regression analysis. In linear trend analysis, the time period is used as the x, the independent or predictor variable, in the analysis to determine the equation of the trend line. In the case of the Huntsville Chemicals Company, the x variable represents the years 2000–2009.

Using sales as the y variable and time (year) as the x variable, the equation of the trend line can be calculated in the usual way as shown in Table 12.9 and is determined to be: $\hat{y} = -5{,}320.56 + 2.6687\,x$. The slope, 2.6687, means that for every yearly increase in time, sales increases by an average of $2.6687 (million). The intercept would represent company sales in the year 0 which, of course, in this problem has no meaning since the Huntsville Chemical Company was not in existence in the year 0. Figure 12.18 is a Minitab display of the Huntsville sales data with the fitted trend line. Note that the output contains the equation of the trend line along with the values of s (standard error of the estimate) and R-Sq (r^2). As is typical with data that have a relatively strong trend, the r^2 value (.963) is quite high.

FIGURE 12.17
Minitab Scatter Plot of Huntsville Chemicals Data

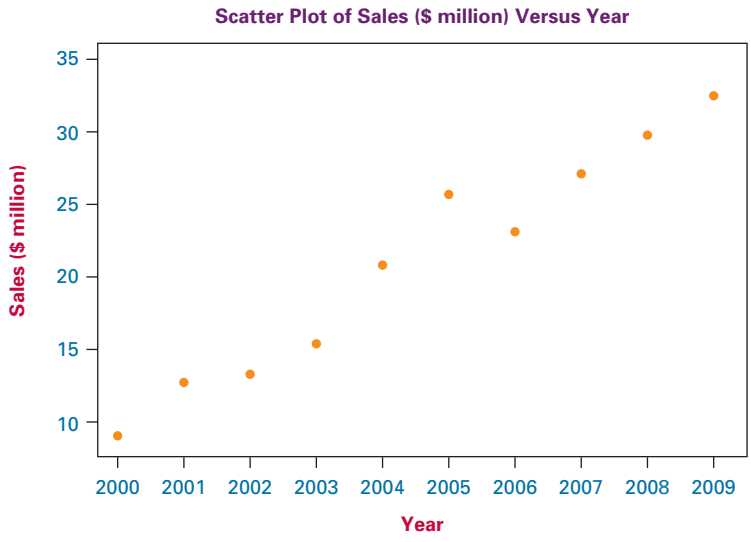


TABLE 12.9

Determining the Equation of the Trend Line for the Huntsville Chemical Company Sales Data

Year x	Sales y	x^2	xy
2000	7.84	4,000,000	15,680.00
2001	12.26	4,004,001	24,532.26
2002	13.11	4,008,004	26,246.22
2003	15.78	4,012,009	31,607.34
2004	21.29	4,016,016	42,665.16
2005	25.68	4,020,025	51,488.40
2006	23.80	4,024,036	47,742.80
2007	26.43	4,028,049	53,045.01
2008	29.16	4,032,064	58,553.28
2009	33.06	4,036,081	66,417.54
$\Sigma x = 20{,}045$	$\Sigma y = 208.41$	$\Sigma x^2 = 40{,}180{,}285$	$\Sigma xy = 417{,}978.01$

$$b_1 = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{(417{,}978.01) - \frac{(20{,}045)(208.41)}{10}}{40{,}180{,}285 - \frac{(20{,}045)^2}{10}} = \frac{220.17}{82.5} = 2.6687$$

$$b_0 = \frac{\sum y}{n} - b_1\frac{\sum x}{n} = \frac{208.41}{10} - (2.6687)\frac{20{,}045}{10} = -5{,}328.57$$

Equation of the Trend Line: $\hat{y} = -5{,}328.57 + 2.6687x$

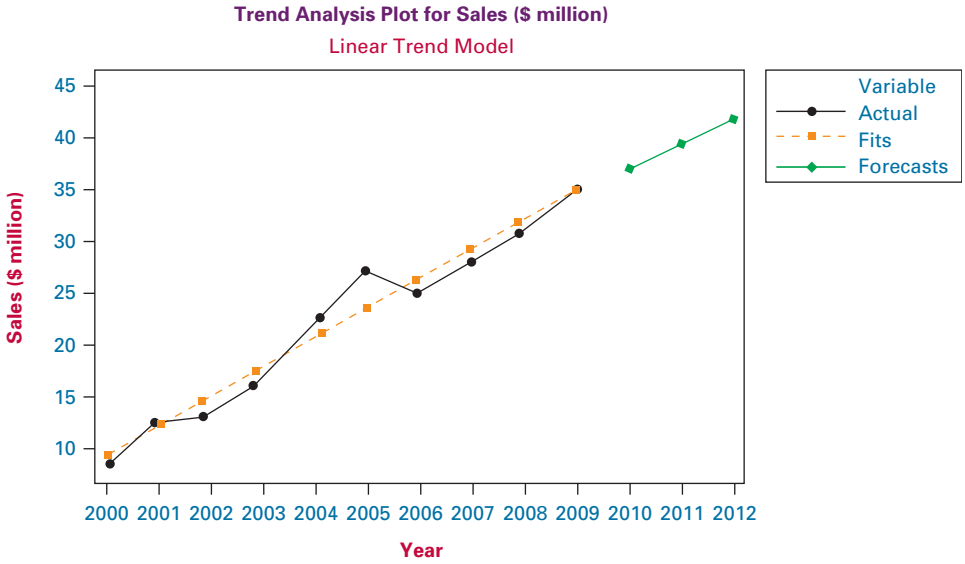
Forecasting Using the Equation of the Trend Line

The main use of the equation of a trend line by business analysts is for forecasting outcomes for time periods in the future. Recall the caution from Section 12.8 that using a regression model to predict y values for x values outside the domain of those used to develop the model may not be valid. Despite this caution and understanding the potential drawbacks, business forecasters nevertheless extrapolate trend lines beyond the most current time periods of the data and attempt to predict outcomes for time periods in the future. To forecast for future time periods using a trend line, insert the time period of interest into the equation of the trend line and solve for $\hat{y}$. For example, suppose forecasters for the Huntsville Chemicals

FIGURE 12.18

Minitab Graph of Huntsville Sales Data with a Fitted Trend Line

FIGURE 12.19
Minitab Output for Trend Line and Forecasts



Company want to predict sales for the year 2012 using the equation of the trend line developed from their historical time series data. Replacing x in the equation of the sales trend line with 2012, results in a forecast of $40.85 (million):

$$\hat{y}(2012) = -5{,}328.57 + 2.6687(2012) = 40.85$$

Figure 12.19 shows Minitab output for the Huntsville Chemicals Company data with the trend line through the data and graphical forecasts for the next three periods (2010, 2011, and 2012). Observe from the graph that the forecast for 2012 is about $41 (million).

Alternate Coding for Time Periods

If you manually calculate the equation of a trend line when the time periods are years, you notice that the calculations can get quite large and cumbersome (observe Table 12.9). However, if the years are consecutive, they can be recoded using many different possible schemes and still produce a meaningful trend line equation (albeit a different y intercept value). For example, instead of using the years 2000–2009, suppose we use the years 1 to 10. That is, 2000 = 1 (first year), 2001 = 2, 2002 = 3, and so on, to 2009 = 10. This recoding scheme produces the trend line equation of: $\hat{y} = 6.1632 + 2.6687x$ as shown in Table 12.10. Notice that the slope of the trend line is the same whether the years 2000 through 2009 are used or the recoded years of 1 through 10, but the y intercept (6.1632) is different. This needs to be taken into consideration when using the equation of the trend line for forecasting. Since the new trend equation was derived from recoded data, forecasts will also need to be made using recoded data. For example, using the recoded system of 1 through 10 to represent "years," the year 2012 is recoded as 13 (2009 = 10, 2010 = 11, 2011 = 12, and 2012 = 13). Inserting this value into the trend line equation results in a forecast of $40.86, the same as the value obtained using raw years as time.

$$\hat{y} = 6.1632 + 2.6687x = 6.1632 + 2.6687(13) = \$40.86 \text{ (million)}.$$

Similar time recoding schemes can be used in the calculating of trend line equations when the time variable is something other than years. For example, in the case of monthly time series data, the time periods can be recoded as:

$$\text{January} = 1, \text{February} = 2, \text{March} = 3, \ldots, \text{December} = 12.$$

TABLE 12.10
Using Recoded Data to Calculate the Trend Line Equation

Year x	Sales y	x^2	xy
1	7.84	1	7.84
2	12.26	4	24.52
3	13.11	9	39.33
4	15.78	16	63.12
5	21.29	25	106.45
6	25.68	36	154.08
7	23.80	49	166.60
8	26.43	64	211.44
9	29.16	81	262.44
10	33.06	100	330.60
$\Sigma x = 55$	$\Sigma y = 208.41$	$\Sigma x^2 = 385$	$\Sigma xy = 1{,}366.42$

$$b_1 = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{(1{,}366.42) - \frac{(55)(208.41)}{10}}{385 - \frac{(55)^2}{10}} = \frac{220.165}{82.5} = 2.6687$$

$$b_0 = \frac{\sum y}{n} - b_1\frac{\sum x}{n} = \frac{208.41}{10} - (2.6687)\frac{55}{10} = 6.1632$$

Equation of the Trend Line: $\hat{y} = 6.1632 + 2.6687x$

In the case of quarterly data over a two-year period, the time periods can be recoded with a scheme such as:

Time Period		Recoded Time Period
Year 1:	Quarter 1	1
	Quarter 2	2
	Quarter 3	3
	Quarter 4	4
Year 2:	Quarter 1	5
	Quarter 2	6
	Quarter 3	7
	Quarter 4	8

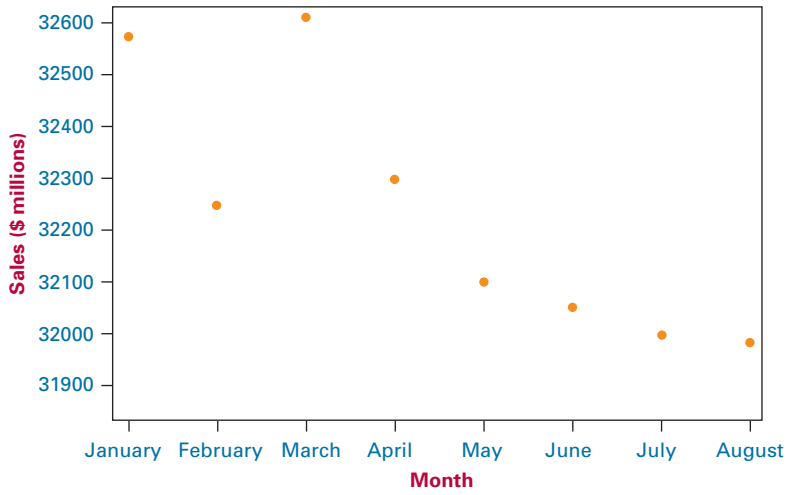
DEMONSTRATION PROBLEM 12.7

Shown below are monthly food and beverage sales in the United States during a recent year over an eight-month period ($ million). Develop the equation of a trend line through these data and use the equation to forecast sales for October.

Month	Sales ($ million)
January	32,569
February	32,274
March	32,583
April	32,304
May	32,149
June	32,077
July	31,989
August	31,977

Solution

Shown here is a Minitab-produced scatter diagram of these time series data:



The months of January through August can be coded using the numbers of 1 through 8, respectively. Using these numbers as the time period values (x) and sales as the dependent variable (y), the following output was obtained from Minitab:

```
Regression Analysis: Sales versus Month

The regression equation is
Sales = 32628 - 86.2 Month
Predictor        Coef     SE Coef          T        P
Constant      32628.2        93.3     349.80    0.000
Month          -86.21       18.47      -4.67    0.003
S = 119.708   R-Sq = 78.4%   R-Sq(adj) = 74.8%
```

The equation of the trend line is: $\hat{y} = 32{,}628.2 - 86.21x$. A slope of –86.21 indicates that there is a downward trend of food and beverage sales over this period of time at a rate of $86.21 (million) per month. The y intercept of 32,628.2 represents what the trend line would estimate the sales to have been in period 0 or December of the previous year. The sales figure for October can be forecast by inserting $x = 10$ into this model and obtaining:

$$\hat{y}(10) = 32{,}628.2 - 86.21(10) = 31{,}766.1.$$

12.9 PROBLEMS

12.48 Determine the equation of the trend line for the data shown below on U.S. exports of fertilizers to Indonesia over a five-year period provided by the U.S Census Bureau. Using the trend line equation, forecast the value for the year 2011.

Year	Fertilizer ($ millions)
2005	11.9
2006	17.9
2007	22.0
2008	21.8
2009	26.0

12.49 Shown below are rental and leasing revenue figures for office machinery and equipment in the United States over a seven-year period according to the U.S. Census Bureau. Use these data to construct a trend line and forecast the rental and leasing revenue for the year 2010 using these data.

Year	Rental and Leasing ($ millions)
2002	5,860
2003	6,632
2004	7,125
2005	6,000
2006	4,380
2007	3,326
2008	2,642

12.50 After a somewhat uncertain start, e-commerce sales in the United States have been growing for the past several years. Shown below are quarterly e-commerce sales figures ($ billions) released by the Census Bureau for the United States over a three-year period. Use these data to determine the equation of a trend line for e-commerce sales during this time and use the trend "model" to forecast e-commerce sales for the third quarter of the year 2010.

Year	Quarter	Sales ($ billions)
2006	1	11.93
	2	12.46
	3	13.28
	4	15.08
2007	1	16.08
	2	16.82
	3	17.60
	4	18.66
2008	1	19.73
	2	21.11
	3	22.21
	4	22.94

12.10 INTERPRETING THE OUTPUT

Although manual computations can be done, most regression problems are analyzed by using a computer. In this section, computer output from both Minitab and Excel will be presented and discussed.

At the top of the Minitab regression output, shown in Figure 12.20, is the regression equation. Next is a table that describes the model in more detail. "Coef" stands for coefficient of the regression terms. The coefficient of Number of Passengers, the x variable, is 0.040702. This value is equal to the slope of the regression line and is reflected in the regression equation. The coefficient shown next to the constant term (1.5698) is the value of the constant, which is the y intercept and also a part of the regression equation. The "T" values are a t test for the slope and a t test for the intercept or constant. (We generally do not interpret the t test for the constant.) The t value for the slope, $t = 9.44$ with an associated probability of .000, is the same as the value obtained manually in section 12.7. Because the probability of the t value is given, the p-value method can be used to interpret the t value.

FIGURE 12.20

Minitab Regression Analysis of the Airline Cost Example

Regression Analysis: Cost ($1,000) versus Number of Passengers

```
The regression equation is
Cost ($1,000) = 1.57 + 0.0407 Number of Passengers

Predictor                Coef   SE Coef      T      P
Constant               1.5698    0.3381   4.64  0.001
Number of Passengers 0.040702  0.004312   9.44  0.000


S = 0.177217 R-Sq = 89.9%    R-Sq(adj) = 88.9%

Analysis of Variance

Source           DF      SS      MS      F      P
Regression        1  2.7980  2.7980  89.09  0.000
Residual Error   10  0.3141  0.0314
Total            11  3.1121

     Number of   Cost
Obs Passengers ($1,000)   Fit   SE Fit  Residual
 1     61.0     4.2800  4.0526  0.0876    0.2274
 2     63.0     4.0800  4.1340  0.0808   -0.0540
 3     67.0     4.4200  4.2968  0.0683    0.1232
 4     69.0     4.1700  4.3782  0.0629   -0.2082
 5     70.0     4.4800  4.4189  0.0605    0.0611
 6     74.0     4.3000  4.5817  0.0533   -0.2817
 7     76.0     4.8200  4.6631  0.0516    0.1569
 8     81.0     4.7000  4.8666  0.0533   -0.1666
 9     86.0     5.1100  5.0701  0.0629    0.0399
10     91.0     5.1300  5.2736  0.0775   -0.1436
11     95.0     5.6400  5.4364  0.0912    0.2036
12     97.0     5.5600  5.5178  0.0984    0.0422
```

The next row of output is the standard error of the estimate s_e, S = 0.177217; the coefficient of determination, r^2, R-Sq = 89.9%; and the adjusted value of r^2, R-Sq(adj) = 88.9%. (Adjusted r^2 will be discussed in Chapter 13.) Following these items is the analysis of variance table. Note that the value of $F = 89.09$ is used to test the overall model of the regression line. The final item of the output is the predicted value and the corresponding residual for each pair of points.

Although the Excel regression output, shown in Figure 12.21 for Demonstration Problem 12.1, is somewhat different from the Minitab output, the same essential regression features are present. The regression equation is found under Coefficients at the bottom of ANOVA. The slope or coefficient of x is 2.2315 and the y-intercept is 30.9125. The standard error of the estimate for the hospital problem is given as the fourth statistic under Regression Statistics at the top of the output, Standard Error = 15.6491. The r^2 value is given as 0.886 on the second line. The t test for the slope is found under t Stat near the bottom of the ANOVA section on the "Number of Beds" (x variable) row, $t = 8.83$. Adjacent to the t Stat is the P-value, which is the probability of the t statistic occurring by chance if the null hypothesis is true. For this slope, the probability shown is 0.000005. The ANOVA table is in the middle of the output with the F value having the same probability as the t statistic, 0.000005, and equaling t^2. The predicted values and the residuals are shown in the Residual Output section.

FIGURE 12.21

Excel Regression Output for Demonstration Problem 12.1

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.942
R Square	0.886
Adjusted R Square	0.875
Standard Error	15.6491
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	19115.0632	19115.0632	78.05	0.000005
Residual	10	2448.9368	244.8937		
Total	11	21564			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	30.9125	13.2542	2.33	0.041888
Number of Beds	2.2315	0.2526	8.83	0.000005

RESIDUAL OUTPUT

Observation	*Predicted FTEs*	*Residuals*
1	82.237	–13.237
2	95.626	–0.626
3	95.626	6.374
4	109.015	8.985
5	124.636	1.364
6	133.562	–8.562
7	142.488	–4.488
8	151.414	26.586
9	173.729	–17.729
10	178.192	5.808
11	200.507	–24.507
12	204.970	20.030

Predicting International Hourly Wages by the Price of a Big Mac

In the Decision Dilemma, questions were raised about the relationship between the price of a Big Mac hamburger and net hourly wages around the world and if a model could be developed to predict net hourly wages by the price of a Big Mac. Data were given for a sample of 27 countries. In exploring the possibility that there is a relationship between these two variables, a Pearson product-moment correlation coefficient, *r*, was computed to be .812. This *r* value indicates that there is a relatively high correlation between the two variables and that developing a regression model to predict one variable by the other has potential. Designating net hourly wages as the *y* or dependent variable and the price of a Big Mac as the *x* or predictor variable, the following regression output was obtained for these data using Excel.

Regression Statistics

Multiple *R*	0.812
R Square	0.660
Adjusted *R* Square	0.646
Standard Error	2.934
Observations	27

ANOVA

	df	**SS**	**MS**	***F***	**Significance *F***
Regression	1	416.929	416.929	48.45	0.0000003
Residual	25	215.142	8.606		
Total	26	632.071			

	Coefficients	Standard Error	*t* Stat	*P*-Value
Intercept	−4.545	1.626	−2.79	0.009828805
Big Mac Price	4.741	0.681	6.96	0.0000003

Taken from this output, the regression model is:

$$\text{Net Hourly Wage} = -4.545 + 4.741\,(\text{Price of Big Mac})$$

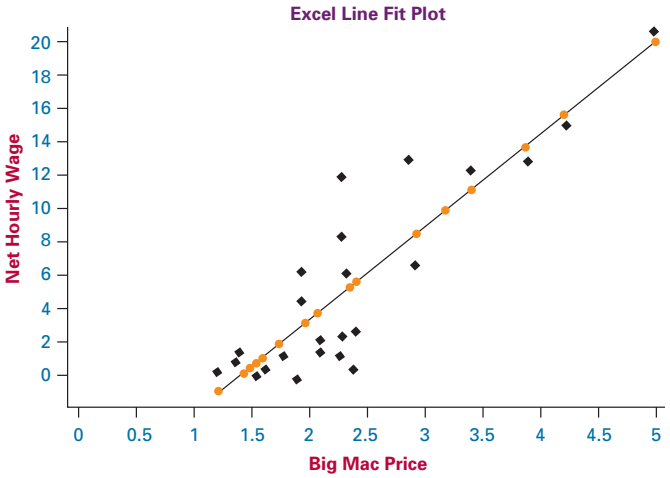
While the *y*-intercept has virtually no practical meaning in this analysis, the slope indicates that for every dollar increase in the price of a Big Mac, there is an incremental increase of $4.741 in net hourly wages for a country. It is worth underscoring here that just because there is a relationship between two variables, it does not mean there is a cause-and-effect relationship. That is, McDonald's cannot raise net hour wages in a country just by increasing the cost of a Big Mac!

Using this regression model, the net hourly wage for a country with a $3.00 Big Mac can be predicted by substituting $x = 3$ into the model:

$$\text{Net Hourly Wage} = -4.545 + 4.741(3) = \$9.68$$

That is, the model predicts that the net hourly wage for a country is $9.68 when the price of a Big Mac is $3.00.

How good a fit is the regression model to the data? Observe from the Excel output that the *F* value for testing the overall significance of the model (48.45) is highly significant with a *p*-value of .0000003, and that the *t* statistic for testing to determine if the slope is significantly different from zero is 6.96 with a *p*-value of .0000003. In simple regression, the *t* statistic is the square root of the *F* value and these statistics relate essentially the same information—that there are significant regression effects in the model. The r^2 value is 66.0%, indicating that the model has moderate predictability. The standard error of the model, $s = 2.934$, indicates that if the error terms are approximately normally distributed, about 68% of the predicted net hourly wages would fall within ± $2.93.



Shown here is an Excel-produced line fit plot. Note from the plot that there generally appears be a linear relationship between the variables but that many of the data points fall considerably away from the fitted regression line, indicating that the price of a Big Mac only partially accounts for net hourly wages.

ETHICAL CONSIDERATIONS

Regression analysis offers several opportunities for unethical behavior. One way is to present a regression model in isolation from information about the fit of the model. That is, the regression model is represented as a valid tool for prediction without any regard for how well it actually fits the data. While it is true that least squares analysis can produce a line of best fit through virtually any set of points, it does not necessarily follow that the regression model is a good predictor of the dependent variable. For example, sometimes business consultants sell regression models to companies as forecasting tools or market predictors without disclosing to the client that the r^2 value is very low, the slope of the regression line is not significant, the residuals are large, and the standard error of the estimate is large. This is unethical behavior.

Another unethical use of simple regression analysis is stating or implying a cause-and-effect relationship between two variables just because they are highly correlated and produce a high r^2 in regression. The Decision Dilemma presents a good example of this with the regression analysis of the price of a Big Mac hamburger and the net hourly wages in a country. While the coefficient of determination is 66.0% and there appears to be a modest fit of the regression line to the data, that does not mean that increasing the price of a Big Mac in a given country will increase the country's net hourly wages. Often, two correlated variables are related to a third variable that drives the two of them but is not included in the regression analysis. In the Decision Dilemma example, both Big Mac prices and net hourly wages may be related to exchange rates or a country's economic condition.

A third way that business analysts can act unethically in using regression analysis is to knowingly violate the assumptions underlying regression. Regression analysis requires equal error variance, independent error terms, and error terms that are normally distributed. Through the use of residual plots and other statistical techniques, a business researcher can test these assumptions. To present a regression model as fact when the assumptions underlying it are being grossly violated is unethical behavior.

It is important to remember that since regression models are developed from sample data, when an x value is entered into a simple regression model, the resulting prediction is only a point estimate. While business people do often use regression models as predicting tools, it should be kept in mind that the prediction value is an estimate not a guaranteed outcome. By utilizing or at least pointing out confidence intervals and prediction intervals, such as those presented in Section 12.8, the business researcher places the predicted point estimate within the context of inferential estimation and is thereby acting more ethically.

And lastly, another ethical problem that arises in regression analysis is using the regression model to predict values of the independent variable that are outside the domain of values used to develop the model. The airline cost model used in this chapter was built with between 61 and 97 passengers. A linear relationship appeared to be evident between flight costs and number of passengers over this domain. This model is not guaranteed to fit values outside the domain of 61 to 97 passengers, however. In fact, either a nonlinear relationship or no relationship may be present between flight costs and number of passengers if values from outside this domain are included in the model-building process. It is a mistake and probably unethical behavior to make claims for a regression model outside the perview of the domain of values for which the model was developed.

SUMMARY

Correlation measures the degree of relatedness of variables. The most well-known measure of correlation is the Pearson product-moment coefficient of correlation, r. This value ranges from –1 to 0 to +1. An r value of +1 is perfect positive correlation and an r value of −1 is perfect negative correlation. Positive correlation means that as one variable increases in value, the other variable tends to increase. Negative correlation means that as one variable increases in value, the other variable tends to decrease. For r values near zero, little or no correlation is present.

Regression is a procedure that produces a mathematical model (function) that can be used to predict one variable by other variables. Simple regression is bivariate (two variables) and linear (only a line fit is attempted). Simple regression analysis produces a model that attempts to predict a y variable, referred to as the dependent variable, by an x variable, referred to as the independent variable. The general form of the equation of the simple regression line is the slope-intercept equation of a line. The equation of the simple regression model consists of a slope of the line as a coefficient of x and a y-intercept value as a constant.

After the equation of the line has been developed, several statistics are available that can be used to determine how well the line fits the data. Using the historical data values of x, predicted values of y (denoted as $\hat{y}$) can be calculated by inserting values of x into the regression equation. The predicted values can then be compared to the actual values of y to determine how well the regression equation fits the known data. The difference between a specific y value and its associated predicted y value is called the residual or error of prediction. Examination of the residuals can offer insight into the magnitude of the errors produced by a model. In addition, residual analysis can be used to help determine whether the assumptions underlying the regression analysis have been met. Specifically, graphs of the residuals can reveal (1) lack of linearity, (2) lack of homogeneity of error variance, and (3) independence of error terms. Geometrically, the residuals are the vertical distances from the y values to the regression line. Because the equation that yields the regression line is derived in such a way that the line is in the geometric middle of the points, the sum of the residuals is zero.

A single value of error measurement called the standard error of the estimate, s_e, can be computed. The standard error of the estimate is the standard deviation of error of a model. The value of s_e can be used as a single guide to the magnitude of the error produced by the regression model as opposed to examining all the residuals.

Another widely used statistic for testing the strength of a regression model is r^2, or the coefficient of determination. The coefficient of determination is the proportion of total variance of the y variable accounted for or predicted by x. The coefficient of determination ranges from 0 to 1. The higher the r^2 is, the stronger is the predictability of the model.

Testing to determine whether the slope of the regression line is different from zero is another way to judge the fit of the regression model to the data. If the population slope of the regression line is not different from zero, the regression model is not adding significant predictability to the dependent variable. A t statistic is used to test the significance of the slope. The overall significance of the regression model can be tested using an F statistic. In simple regression, because only one predictor is present, this test accomplishes the same thing as the t test of the slope and $F = t^2$.

One of the most prevalent uses of a regression model is to predict the values of y for given values of x. Recognizing that the predicted value is often not the same as the actual value, a confidence interval has been developed to yield a range within which the mean y value for a given x should fall. A prediction interval for a single y value for a given x value also is specified. This second interval is wider because it allows for the wide diversity of individual values, whereas the confidence interval for the mean y value reflects only the range of average y values for a given x.

Time-series data are data that are gathered over a period of time at regular intervals. Developing the equation of a forecasting trend line for time-series data is a special case of simple regression analysis where the time factor is the predictor variable. The time variable can be in units of years, months, weeks, quarters, and others.

KEY TERMS

coefficient of determination (r^2)
confidence interval
dependent variable
deterministic model
heteroscedasticity
homoscedasticity
independent variable
least squares analysis
outliers
prediction interval
probabilistic model
regression analysis
residual
residual plot
scatter plot
simple regression
standard error of the estimate (s_e)
sum of squares of error (SSE)

FORMULAS

Pearson's product-moment correlation coefficient

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

Equation of the simple regression line

$$\hat{y} = \beta_0 + \beta_1 x$$

Sum of squares

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

$$SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

$$SS_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n}$$

Slope of the regression line

$$b_1 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2} = \frac{\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}$$

y-intercept of the regression line

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{\Sigma y}{n} - b_1 \frac{(\Sigma x)}{n}$$

Sum of squares of error

$$SSE = \Sigma(y - \hat{y})^2 = \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy$$

Standard error of the estimate

$$s_e = \sqrt{\frac{SSE}{n - 2}}$$

Coefficient of determination

$$r^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}}$$

Computational formula for r^2

$$r^2 = \frac{b_1^2 SS_{xx}}{SS_{yy}}$$

t test of slope

$$t = \frac{b_1 - \beta_1}{s_b}$$

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}}$$

Confidence interval to estimate $E(y_x)$ for a given value of x

$$\hat{y} \pm t_{\alpha/2,\, n-2} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

Prediction interval to estimate y for a given value of x

$$\hat{y} \pm t_{\alpha/2,\, n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

SUPPLEMENTARY PROBLEMS

CALCULATING THE STATISTICS

12.51 Determine the Pearson product-moment correlation coefficient for the following data.

x	1	10	9	6	5	3	2
y	8	4	4	5	7	7	9

12.52 Use the following data for parts (a) through (f).

x	5	7	3	16	12	9
y	8	9	11	27	15	13

a. Determine the equation of the least squares regression line to predict y by x.

b. Using the x values, solve for the predicted values of y and the residuals.
c. Solve for s_e.
d. Solve for r^2.
e. Test the slope of the regression line. Use $\alpha = .01$.
f. Comment on the results determined in parts (b) through (e), and make a statement about the fit of the line.

12.53 Use the following data for parts (a) through (g).

x	53	47	41	50	58	62	45	60
y	5	5	7	4	10	12	3	11

a. Determine the equation of the simple regression line to predict y from x.
b. Using the x values, solve for the predicted values of y and the residuals.
c. Solve for SSE.
d. Calculate the standard error of the estimate.
e. Determine the coefficient of determination.
f. Test the slope of the regression line. Assume $\alpha = .05$. What do you conclude about the slope?
g. Comment on parts (d) and (e).

12.54 If you were to develop a regression line to predict y by x, what value would the coefficient of determination have?

x	213	196	184	202	221	247
y	76	65	62	68	71	75

12.55 Determine the equation of the least squares regression line to predict y from the following data.

x	47	94	68	73	80	49	52	61
y	14	40	34	31	36	19	20	21

a. Construct a 95% confidence interval to estimate the mean y value for $x = 60$.
b. Construct a 95% prediction interval to estimate an individual y value for $x = 70$.
c. Interpret the results obtained in parts (a) and (b).

12.56 Determine the equation of the trend line through the following cost data. Use the equation of the line to forecast cost for year 7.

Year	Cost ($ millions)
1	56
2	54
3	49
4	46
5	45

TESTING YOUR UNDERSTANDING

12.57 A manager of a car dealership believes there is a relationship between the number of salespeople on duty and the number of cars sold. Suppose the following sample is used to develop a simple regression model to predict the number of cars sold by the number of salespeople. Solve for r^2 and explain what r^2 means in this problem.

Week	Number of Cars Sold	Number of Salespeople
1	79	6
2	64	6
3	49	4
4	23	2
5	52	3

12.58 Executives of a video rental chain want to predict the success of a potential new store. The company's researcher begins by gathering information on number of rentals and average family income from several of the chain's present outlets.

Rentals	Average Family Income ($1,000)
710	65
529	43
314	29
504	47
619	52
428	50
317	46
205	29
468	31
545	43
607	49
694	64

Develop a regression model to predict the number of rentals per day by the average family income. Comment on the output.

12.59 It seems logical that restaurant chains with more units (restaurants) would have greater sales. This assumption is mitigated, however, by several possibilities: some units may be more profitable than others, some units may be larger, some units may serve more meals, some units may serve more expensive meals, and so on. The data shown here were published by Technomic. Perform a simple regression analysis to predict a restaurant chain's sales by its number of units. How strong is the relationship?

Chain	Sales ($ billions)	Number of Units (1000)
McDonald's	17.1	12.4
Burger King	7.9	7.5
Taco Bell	4.8	6.8
Pizza Hut	4.7	8.7
Wendy's	4.6	4.6
KFC	4.0	5.1
Subway	2.9	11.2
Dairy Queen	2.7	5.1
Hardee's	2.7	2.9

12.60 Shown here are the total employment labor force figures for the country of Romania over a 13-year period

published in LABORSTA. Develop the equation of a trend line through these data and use the equation to predict the total employment labor force of Romania for the year 2011.

Year	Total Employment (1000s)
1995	11,152
1996	10,935
1997	11,050
1998	10,845
1999	10,776
2000	10,764
2001	10,697
2002	9,234
2003	9,223
2004	9,158
2005	9,147
2006	9,313
2007	9,353

12.61 How strong is the correlation between the inflation rate and 30-year treasury yields? The following data published by Fuji Securities are given as pairs of inflation rates and treasury yields for selected years over a 35-year period.

Inflation Rate	30-Year Treasure Yield
1.57%	3.05%
2.23	3.93
2.17	4.68
4.53	6.57
7.25	8.27
9.25	12.01
5.00	10.27
4.62	8.45

Compute the Pearson product-moment correlation coefficient to determine the strength of the correlation between these two variables. Comment on the strength and direction of the correlation.

12.62 According to the National Marine Fisheries Service, the current landings in millions of pounds of fish by U.S. fleets are almost double what they were in the 1970s. In other words, fishing has not faded as an industry. However, the growth of this industry has varied by region as shown in the following data. Some regions have remained relatively constant, the South Atlantic region has dropped in pounds caught, and the Pacific-Alaska region has grown more than threefold.

Fisheries	1977	2006
New England	581	752
Mid-Atlantic	213	213
Chesapeake	668	477
South Atlantic	345	114
Gulf of Mexico	1476	1286
Pacific-Alaska	1776	6602

Develop a simple regression model to predict the 2006 landings by the 1977 landings. According to the model, if a region had 700 landings in 1977, what would the predicted number be for 2006? Construct a confidence interval for the average y value for the 700 landings. Use the t statistic to test to determine whether the slope is significantly different from zero. Use $\alpha = .05$.

12.63 People in the aerospace industry believe the cost of a space project is a function of the weight of the major object being sent into space. Use the following data to develop a regression model to predict the cost of a space project by the weight of the space object. Determine r^2 and s_e.

Weight (tons)	Cost ($ millions)
1.897	$ 53.6
3.019	184.9
0.453	6.4
0.988	23.5
1.058	33.4
2.100	110.4
2.387	104.6

12.64 The following data represent a breakdown of state banks and all savings organizations in the United States every 5 years over a 60-year span according to the Federal Reserve System.

Time Period	State Banks	All Savings
1	1342	2330
2	1864	2667
3	1912	3054
4	1847	3764
5	1641	4423
6	1405	4837
7	1147	4694
8	1046	4407
9	997	4328
10	1070	3626
11	1009	2815
12	1042	2030
13	992	1779

Develop a regression model to predict the total number of state banks by the number of all savings organizations. Comment on the strength of the model. Develop a time-series trend line for All Savings using the time periods given. Forecast All Savings for period 15 using this equation.

12.65 Is the amount of money spent by companies on advertising a function of the total sales of the company?

Show are sales income and advertising cost data for seven companies published by *Advertising Age*.

Company	Advertising ($ millions)	Sales ($ billions)
Wal-Mart	1,073	351.1
Procter & Gamble	4,898	68.2
AT&T	3,345	63.1
General Motors	3,296	207.3
Verizon	2,822	93.2
Ford Motor	2,577	160.1
Hewlett-Packard	829	91.7

Use the data to develop a regression line to predict the amount of advertising by sales. Compute s_e and r^2. Assuming $\alpha = .05$, test the slope of the regression line. Comment on the strength of the regression model.

12.66 Can the consumption of water in a city be predicted by temperature? The following data represent a sample of a day's water consumption and the high temperature for that day.

Water Use (millions of gallons)	Temperature (degrees Fahrenheit)
219	103°
56	39
107	77
129	78
68	50
184	96
150	90
112	75

Develop a least squares regression line to predict the amount of water used in a day in a city by the high temperature for that day. What would be the predicted water usage for a temperature of 100°? Evaluate the regression model by calculating s_e, by calculating r^2, and by testing the slope. Let $\alpha = .01$.

INTERPRETING THE OUTPUT

12.67 Study the following Minitab output from a regression analysis to predict y from x.

a. What is the equation of the regression model?
b. What is the meaning of the coefficient of x?
c. What is the result of the test of the slope of the regression model? Let $\alpha = .10$. Why is the t ratio negative?
d. Comment on r^2 and the standard error of the estimate.
e. Comment on the relationship of the F value to the t ratio for x.
f. The correlation coefficient for these two variables is −.7918. Is this result surprising to you? Why or why not?

```
Regression Analysis: Y versus X

The regression equation is
Y = 67.2 - 0.0565 X

Predictor   Coef     SE Coef    T       p
Constant  67.231     5.046    13.32   0.000
X        -0.05650   0.01027   -5.50   0.000

S = 10.32   R-Sq = 62.7%   R-Sq(adj) = 60.6%

Analysis of Variance
Source          DF    SS       MS     F      P
Regression       1  3222.9   3222.9 30.25  0.000
Residual Error  18  1918.0    106.6
Total           19  5141.0
```

12.68 Study the following Excel regression output for an analysis attempting to predict the number of union members in the United States by the size of the labor force for selected years over a 30-year period from data published by the U.S. Bureau of Labor Statistics. Analyze the computer output. Discuss the strength of the model in terms of proportion of variation accounted for, slope, and overall predictability. Using the equation of the regression line, attempt to predict the number of union members when the labor force is 110,000. Note that the model was developed with data already recoded in 1,000 units. Use the data in the model as is.

SUMMARY OUTPUT

Regression Statistics

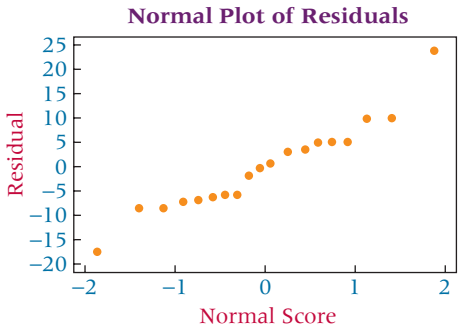
Multiple R	0.798
R Square	0.636
Adjusted R Square	0.612
Standard Error	258.632
Observations	17

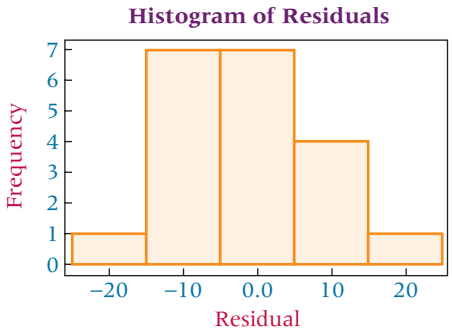
ANOVA

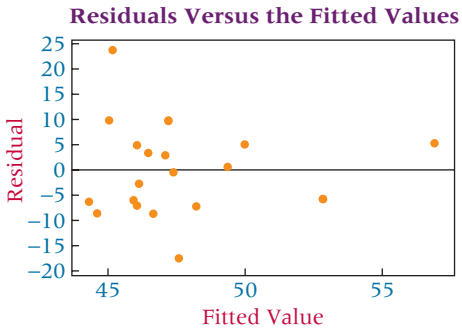
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1756035.529	1756036	26.25	0.00012
Residual	15	1003354.471	66890.3		
Total	16	2759390			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	20704.3805	879.6067	23.54	0.00000
Total Employment	–0.0390	0.0076	–5.12	0.00012

12.69 Study the following Minitab residual diagnostic graphs. Comment on any possible violations of regression assumptions.







ANALYZING THE DATABASES

see www.wiley.com/college/black and WileyPLUS

1. Develop a regression model from the Consumer Food database to predict Annual Food Spending by Annual Household Income. Discuss the model and its strength on the basis of statistics presented in this chapter. Now develop a regression model to predict Non-Mortgage Household Debt by Annual Household Income. Discuss this model and its strengths. Compare the two models. Does it make sense that Annual Food Spending and Non-Mortgage Household Debt could each be predicted by Annual Household Income? Why or why not?
2. Using the Hospital database, develop a regression model to predict the number of Personnel by the number of Births. Now develop a regression model to predict number of Personnel by number of Beds. Examine the regression output. Which model is stronger in predicting number of Personnel? Explain why, using techniques presented in this chapter. Use the second regression model to predict the number of Personnel in a hospital that has 110 beds. Construct a 95% confidence interval around this prediction for the average value of y.
3. Analyze all the variables except Type in the Financial database by using a correlation matrix. The seven variables in this database are capable of producing 21 pairs of correlations. Which are most highly correlated? Select the variable that is most highly correlated with P/E ratio and use it as a predictor to develop a regression model to predict P/E ratio. How did the model do?
4. Construct a correlation matrix for the six U.S. and international stock indicators. Describe what you find. That is, what indicators seem to be most strongly related to other indicators? Now focus on the three international stock indicators. Which pair of these stock indicators is most correlated? Develop a regression model to predict the DJIA by the Nikkei 225. How strong is the model? Develop a regression model to predict the DJIA by the Hang Seng. How strong is the model? Develop a regression model to predict the DJIA by the Mexico IPC. How strong is the model? Compare the three models.

CASE

DELTA WIRE USES TRAINING AS A WEAPON

The Delta Wire Corporation was founded in 1978 in Clarksdale, Mississippi. The company manufactures high-carbon specialty steel wire for global markets and at present employs about 100 people. For the past few years, sales increased each year.

A few years ago, however, things did not look as bright for Delta Wire because it was caught in a potentially disastrous bind. With the dollar declining in value, foreign competition was becoming a growing threat to Delta's market position. In addition to the growing foreign competition, industry quality requirements were becoming tougher each year.

Delta officials realized that some conditions, such as the value of the dollar, were beyond their control. However, one area that they could improve upon was employee education. The company worked with training programs developed by the state of Mississippi and a local community college to set up its own school. Delta employees were introduced to statistical process control and other quality assurance techniques. Delta reassured its customers that the company was working hard on improving quality and staying competitive. Customers were invited to sit in on the educational sessions. Because of this effort, Delta has been able to weather the storm and continues to sustain a leadership position in the highly competitive steel wire industry.

Delta continued its training and education program. In the 1990s, Delta instituted a basic skills training program

that eventually led to a decrease in nonconforming material from 6% to 2% and a productivity increase from 70,000 to 90,000 pounds per week. In addition, this initiative resulted in a "best in class" award from Goodyear, its largest customer.

Although acquired by Bekaert of Belgium in January of 2006, the Delta Wire Corporation, a major supplier of bead wire for tire reinforcement and other specialized wire products for the North American market, continues to operate in its current capacity. Bekaert wants to support Delta Wire's market share growth and ensure adequate product availability to its customers.

Discussion

1. Delta Wire prides itself on its efforts in the area of employee education. Employee education can pay off in many ways. Discuss some of them. One payoff can be the renewed interest and excitement generated toward the job and the company. Some people theorize that because of a more positive outlook and interest in implementing things learned, the more education received by a worker, the less likely he or she is to miss work days. Suppose the following data represent the number of days of sick leave taken by 20 workers last year along with the number of contact hours of employee education/training they each received in the past year. Use the techniques learned in this chapter to analyze the data. Include both regression and correlation techniques. Discuss the strength of the relationship and any models that are developed.

Employee	Hours of Education	Sick Days	Employee	Hours of Education	Sick Days
1	24	5	11	8	8
2	16	4	12	60	1
3	48	0	13	0	9
4	120	1	14	28	3
5	36	5	15	15	8
6	10	7	16	88	2
7	65	0	17	120	1
8	36	3	18	15	8
9	0	12	19	48	0
10	12	8	20	5	10

2. Many companies find that the implementation of total quality management eventually results in improved sales. Companies that fail to adopt quality efforts lose market share in many cases or go out of business. One measure of the effect of a company's quality improvement efforts is customer satisfaction. Suppose Delta Wire hired a research firm to measure customer satisfaction each year. The research firm developed a customer satisfaction scale in which totally satisfied customers can award a score as high as 50 and totally unsatisfied customers can award scores as low as 0. The scores are measured across many different industrial customers and averaged for a yearly mean customer score. Do sales increase with increases in customer satisfaction scores? To study this notion, suppose the average customer satisfaction score each year for Delta Wire is paired with the company's total sales of that year for the last 15 years, and a regression analysis is run on the data. Assume the following Minitab and Excel outputs are the result. Suppose you were asked by Delta Wire to analyze the data and summarize the results. What would you find?

MINITAB OUTPUT

```
Regression Analysis: Sales Versus Satisfaction

The regression equation is
Sales = 1.73 + 0.162 CustSat

Predictor      Coef    StDev       T      p
Constant     1.7332   0.4364    3.97  0.002
CustSat     0.16245  0.01490   10.90  0.000

S = 0.4113   R-Sq = 90.1% R-Sq(adj) = 89.4%

Analysis of Variance

Source          DF  SS      MS      F       p
Regression       1  20.098  20.098  118.80  0.000
Residual Error  13   2.199   0.169
Total           14  22.297
```

EXCEL OUTPUT

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.949
R Square	0.901
Adjusted R Square	0.894
Standard Error	0.411
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	20.098	20.098	118.8	0.000
Residual	13	2.199	0.169		
Total	14	22.297			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1.733	0.436	3.97	0.0016
Sick Days	0.162	0.015	10.90	0.0000

3. Delta Wire increased productivity from 70,000 to 90,000 pounds per week during a time when it instituted a basic skills training program. Suppose this program was implemented over an 18-month period and that the following data are the number of total cumulative basic skills hours of training and the per week productivity figures taken once a month over this time. Use techniques from this chapter to analyze the data and make a brief report to Delta about the predictability of productivity from cumulative hours of training.

Cumulative Hours of Training	Productivity (in pounds per week)
0	70,000
100	70,350
250	70,500
375	72,600
525	74,000
750	76,500
875	77,000
1,100	77,400
1,300	77,900
1,450	77,200
1,660	78,900
1,900	81,000
2,300	82,500
2,600	84,000
2,850	86,500
3,150	87,000
3,500	88,600
4,000	90,000

Source: Adapted from "Delta Wire Corporation," *Strengthening America's Competitiveness: Resource Management Insights for Small Business Success.* Published by Warner Books on behalf of Connecticut Mutual Life Insurance Company and the U.S. Chamber of Commerce in association with the Blue Chip Enterprise Initiative, 1991, International Monetary Fund; Terri Bergman, "TRAINING: The Case for Increased Investment," *Employment Relations Today*, Winter 1994–1995, pp. 381–391, available at http://www.ed.psu.edu/nwac/document/train/invest.html. Bekaert Web site at: http://www.bekaert.com/corporate/press/2006/31-jan-2006.htm.

(*continued*)

USING THE COMPUTER

EXCEL

- Excel has the capability of doing simple regression analysis. For a more inclusive analysis, use the **Data Analysis** tool. For a more "a la carte" approach, use Excel's **Insert Function**.
- To use the **Data Analysis** tool for a more inclusive analysis, begin by selecting the **Data** tab on the Excel worksheet. From the **Analysis** panel at the right top of the **Data** tab worksheet, click on **Data Analysis**. If your Excel worksheet does not show the **Data Analysis** option, then you can load it as an add-in following directions given in Chapter 2. From the **Data Analysis** pulldown menu, select **Regression**. In the **Regression** dialog box, input the location of the y values in **Input Y Range**. Input the location of the x values in **Input X Range**. Input **Labels** and input **Confidence Level**. To pass the line through the origin, check **Constant is Zero**. To print out the raw residuals, check **Residuals**. To print out residuals converted to z scores, check **Standardized Residuals**. For a plot of the residuals, check **Residual Plots**. For a plot of the line through the points, check **Line Fit Plots**. Standard output includes r, r^2, s_e, and an ANOVA table with the F test, the slope and intercept, t statistics with associated p-values, and any optionally requested output such as graphs or residuals.
- To use the **Insert Function** (f_x) go to the **Formulas** tab on an Excel worksheet (top center tab). The **Insert Function** is on the far left of the menu bar. In the **Insert Function** dialog box at the top, there is a pulldown menu where it says **Or select a category**. From the pulldown menu associated with this command, select **Statistical**. Select **INTERCEPT** from the **Insert Function's Statistical** menu to solve for the y-intercept, **RSQ** to solve for r^2, **SLOPE** to solve for the slope, and **STEYX** to solve for the standard error of the estimate.

MINITAB

- Minitab has a relatively thorough capability to perform regression analysis. To begin, select **Stat** from the menu bar. Select **Regression** from the **Stat** pulldown menu. Select **Regression** from the **Regression** pulldown menu. Place the column name or column location of the y variable in **Response**. Place the column name or column location of the x variable in **Predictors**. Select **Graphs** for options relating to residual plots. Use this option and check **Four in one** to produce the residual diagnostic plots shown in the chapter. Select **Options** for confidence intervals and prediction intervals. Select **Results** for controlling the regression analysis output. Select **Storage** to store fits and/or residuals.
- To obtain a fitted-line plot, select **Stat** from the menu bar. Select **Regression** from the **Stat** pulldown menu. Select **Fitted Line Plot** from the **Regression** pulldown menu. In the Fitted Line Plot dialog box, place the column name or column location of the y variable in **Response(Y).**
- Place the column name or column location of the x variable in **Response(X)**.Check **Type of Regression Model** as **Linear** (Chapter 12), **Quadratic**,or **Cubic**.
- Select **Graphs** for options relating to residual plots. Use this option and check **Four in one** to produce the residual diagnostic plots shown in the chapter.
- Select **Options** for confidence intervals and prediction intervals.
- Select **Storage** to store fits and/or residuals.