12.3. In an effort to determine whether any correlation exists between the price of stocks of banks, an analyst sampled six days of activity of the stock market spread out over one month. Using the following prices of SBI stock and HDFC stock, compute the coefficient of correlation.

SBI	HDFC
422.70	1581.40
426.05	1568.25
412.45	1548.45
410.75	1554.80
416.25	1558.85
417.60	1524.60

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	$1425	$277
Colorado	273	100
Florida	915	120
Illinois	1687	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 Linear Regression

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 ($1,000s)
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

For example, 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 Minitab. Try to imagine a line passing through the points. Is a linear fit possible? Would a curve fit the data

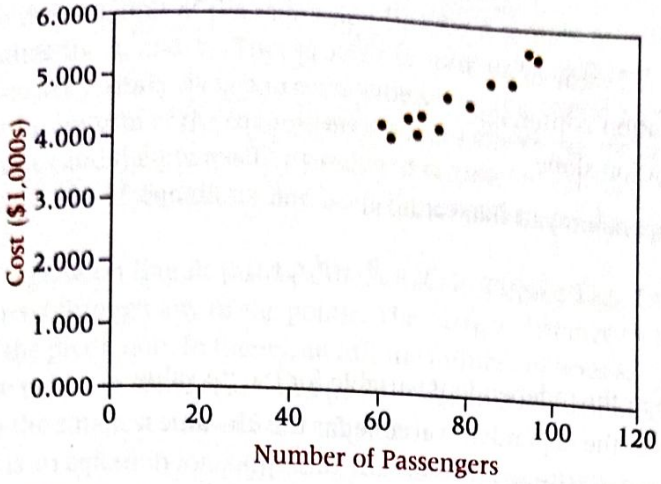


FIGURE 12.3 **Excel Scatter Plot of Airline Cost Data**

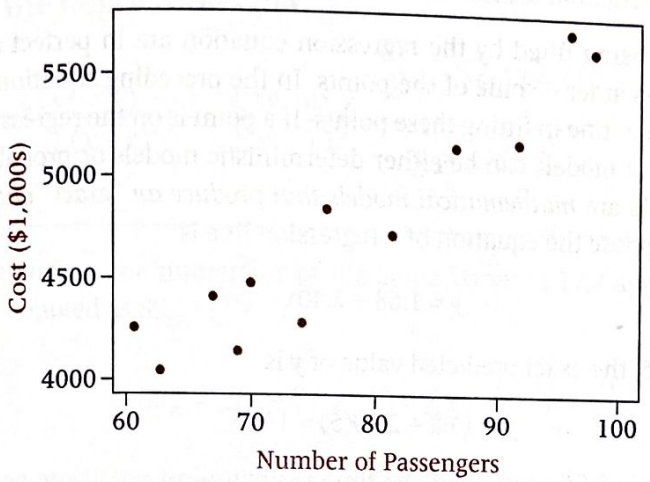


FIGURE 12.4 **Close-Up Minitab Scatter Plot of Airline Cost Data**

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

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 various math 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, business analysts 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 data 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,

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

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
ϵ_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, ϵ_i represents the error of the regression line in fitting these points. If a point is on the regression line, $\epsilon_i = 0$.

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 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

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

$\beta_0 + \beta_1 x$ is the deterministic portion of the probabilistic model, $\beta_0 + \beta_1 x + \epsilon$. In a deterministic model, all points are assumed to be on the line and in all cases ϵ 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 y-intercept
b_1 = the sample slope

To determine the equation of the regression line for a sample of data, the analyst 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 where the errors are the differences between actual values and their respective predicted values. On the basis of this premise and calculus, a particular set of equations has been developed to produce components of the regression model.*

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

$$b_1 = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{\Sigma x^2 - n\bar{x}^2} = \frac{\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}} \qquad (12.2)$$

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

$$b_1 = \frac{SS_{xy}}{SS_{xx}} \qquad (12.3)$$

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

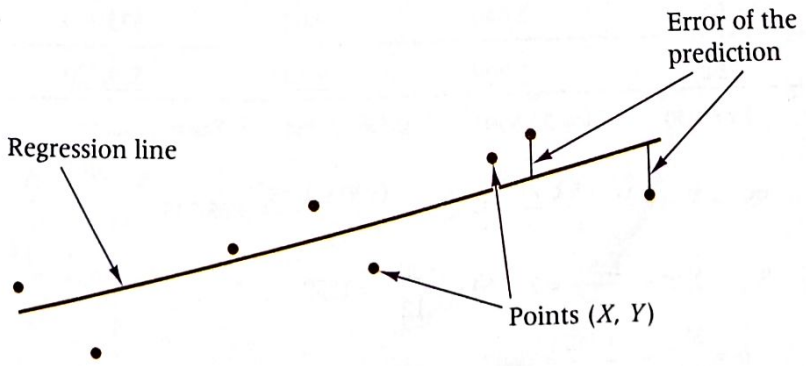


FIGURE 12.5 **Minitab Plot of a Regression Line**

*Derivation of these formulas is beyond our scope here but is presented in WileyPLUS.

y-Intercept of the Regression Line

$$b_0 = \bar{y} - b_1\bar{x} = \frac{\Sigma y}{n} - b_1\frac{(\Sigma x)}{n} \tag{12.4}$$

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

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

The slope of this regression line is .0407. Because the x values were recoded for the ease of computation and are 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.

TABLE 12.4 Solving for the Slope and the y-Intercept of the Regression Line for the Airline Cost Example

Number of Passengers	Cost (\$1,000s)		
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
$\Sigma x = 930$	$\Sigma y = 56.690$	$\Sigma x^2 = 73{,}764$	$\Sigma xy = 4462.220$

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

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

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

$$b_0 = \frac{\Sigma y}{n} - b_1\frac{\Sigma x}{n} = \frac{56.69}{12} - (.0407)\frac{930}{12} = 1.57$$

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

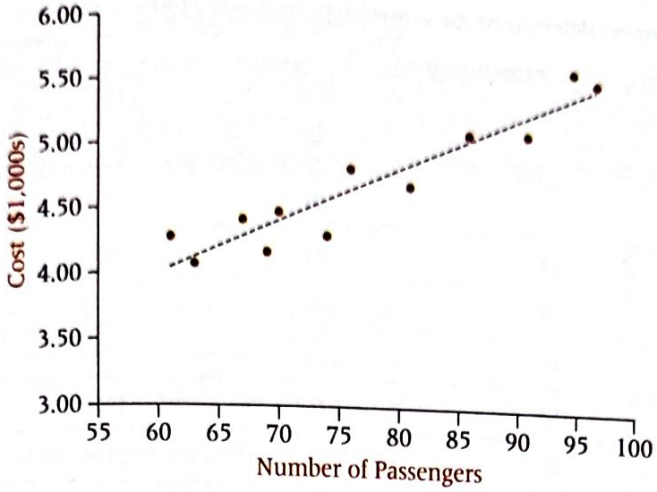


FIGURE 12.6 Excel Graph of Regression Line for the Airline Cost Example

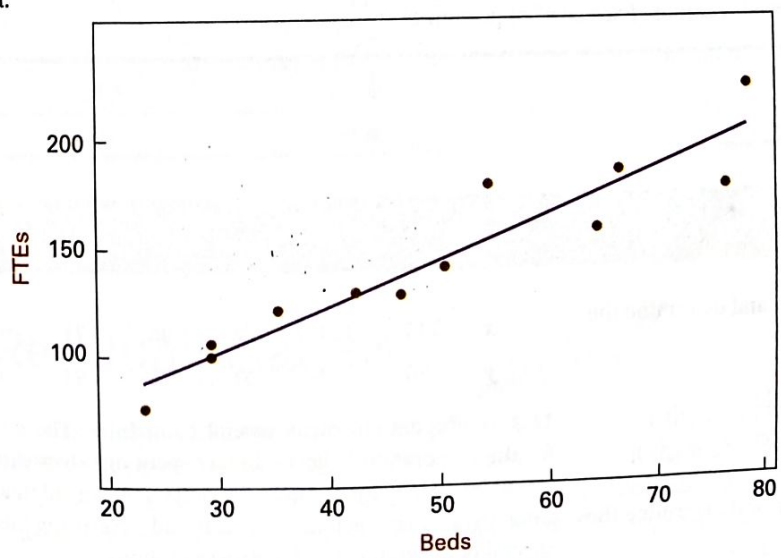
Superimposing the line representing the least squares equation for this problem on the scatter plot indicates how well the regression line fits the data points, as shown in the Excel graph in **Figure 12.6**. The next several sections explore mathematical ways of testing how well the regression line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

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 analyst 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	1587
2	29	95	841	2755
3	29	102	841	2958
4	35	118	1225	4130
5	42	126	1764	5292
6	46	125	2116	5750
7	50	138	2500	6900
8	54	178	2916	9612
9	64	156	4096	9984
10	66	184	4356	12,144
11	76	176	5776	13,376
12	78	225	6084	17,550
	$\Sigma x = 592$	$\Sigma y = 1692$	$\Sigma x^2 = 33,044$	$\Sigma xy = 92,038$

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

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

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

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

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

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 additional bed), y (number of FTEs) is predicted to increase by 2.232. Even though the y-intercept helps the analyst 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	14	23	30	10	22
y	20	18	25	22	27

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

x	145	124	108	96	71	34	29
y	30	34	51	75	93	117	133

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 2019 ($ 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 %
4	15
11	5
8	7
14	3
6	6

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 (1000s)	Firm Births (10,000s)
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 50 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)
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
2010	2.19	419
2015	2.08	438

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 below 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 analyst 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.* A residual is geometrically the vertical distance from the regression line to a data point. The equations used to solve for the slope 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 business analyst 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 analyst 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.

TABLE 12.5 Predicted Values and Residuals for the Airline Cost Example

Number of Passengers x	Cost ($1,000s) 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
			$\Sigma(y - \hat{y}) = -.001$

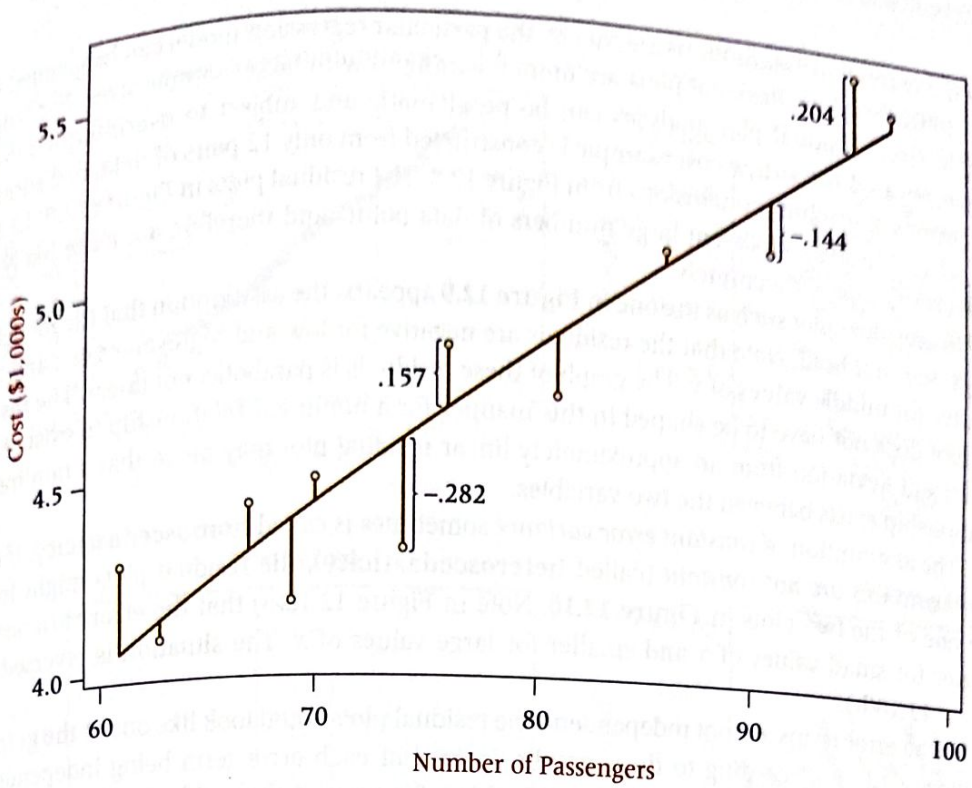


FIGURE 12.7 **Close-Up Minitab Scatter Plot with Residuals for the Airline Cost Example**

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.

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

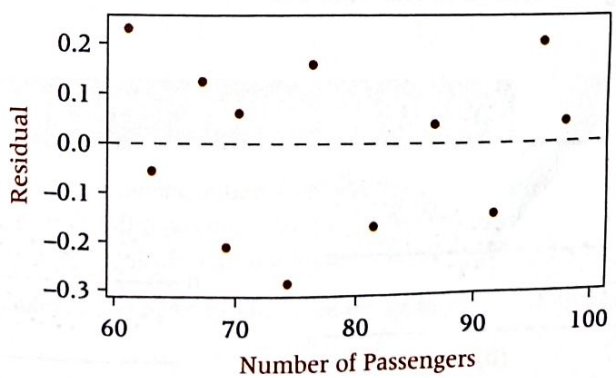


FIGURE 12.8 **Excel Graph of Residuals for the Airline Cost Example**

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 overinterpretation. 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 with a zero mean; 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.

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).

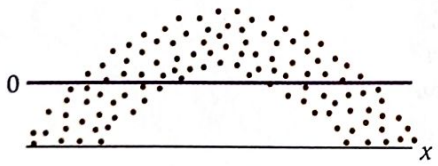


FIGURE 12.9 Nonlinear Residual Plot

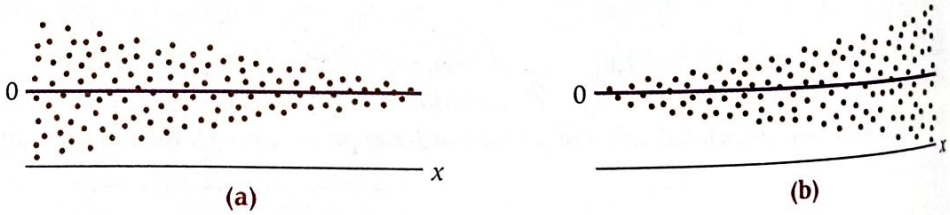
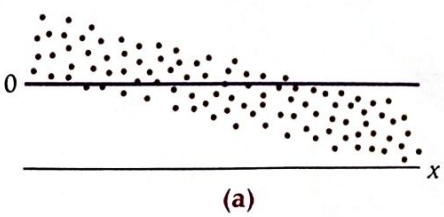


FIGURE 12.10 Nonconstant Error Variance



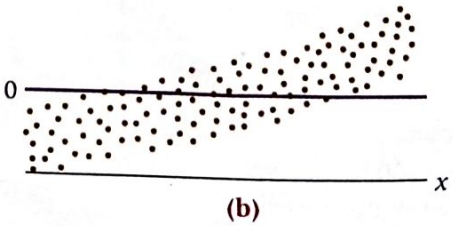


FIGURE 12.11 Graphs of Nonindependent Error Terms

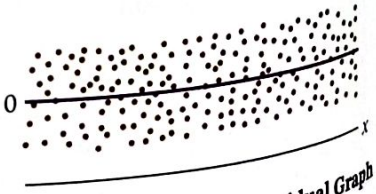


FIGURE 12.12 Healthy Residual Graph

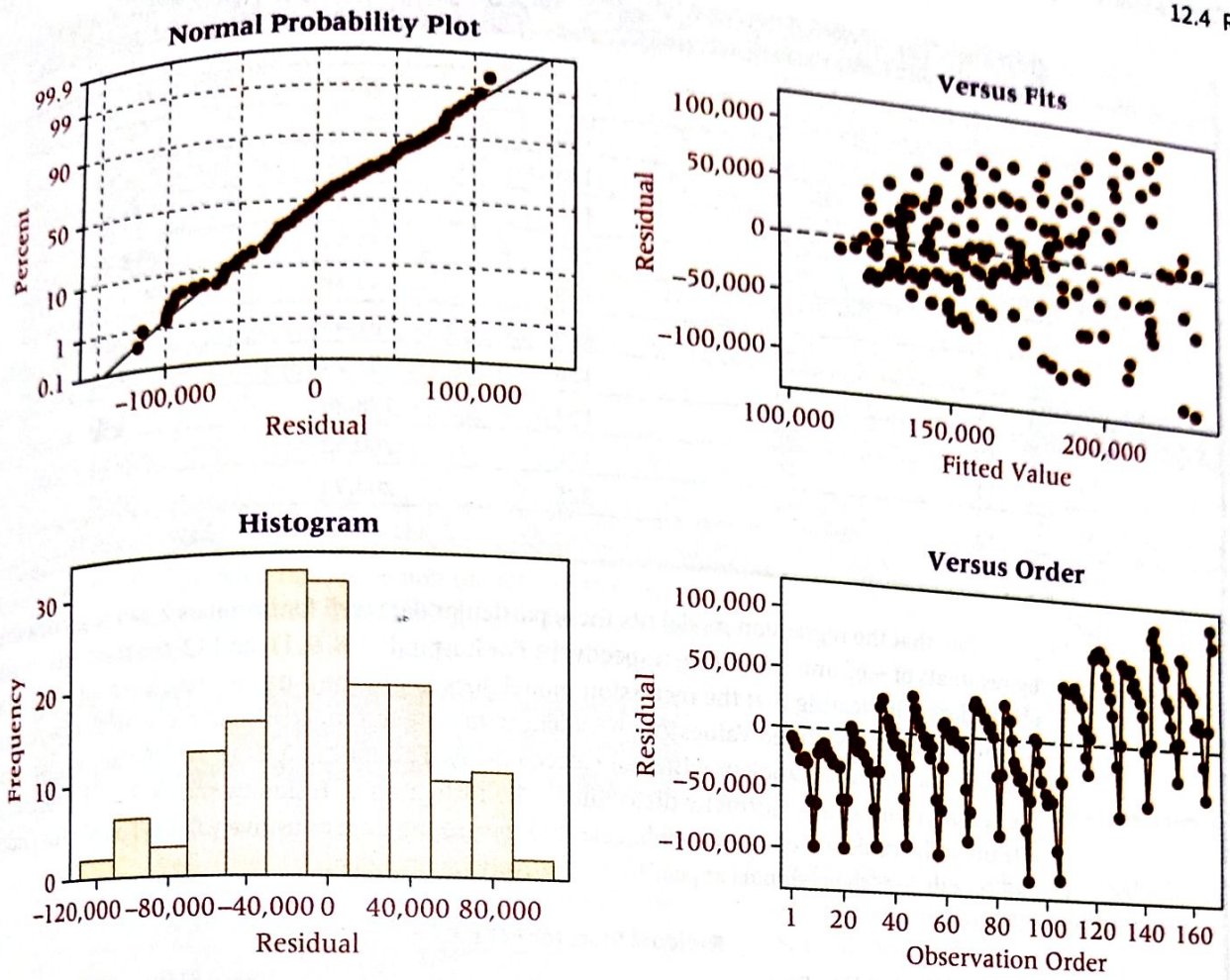


FIGURE 12.13 **Minitab Residual Analyses**

These Minitab residual model diagnostics consist of four 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 analyst 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. The bottom right graph, Versus Order, can be used to verify that the error terms are independent. In this particular graph, note that many of the points seem to follow a pattern indicating that the residuals may be correlated and, thus, not independent.

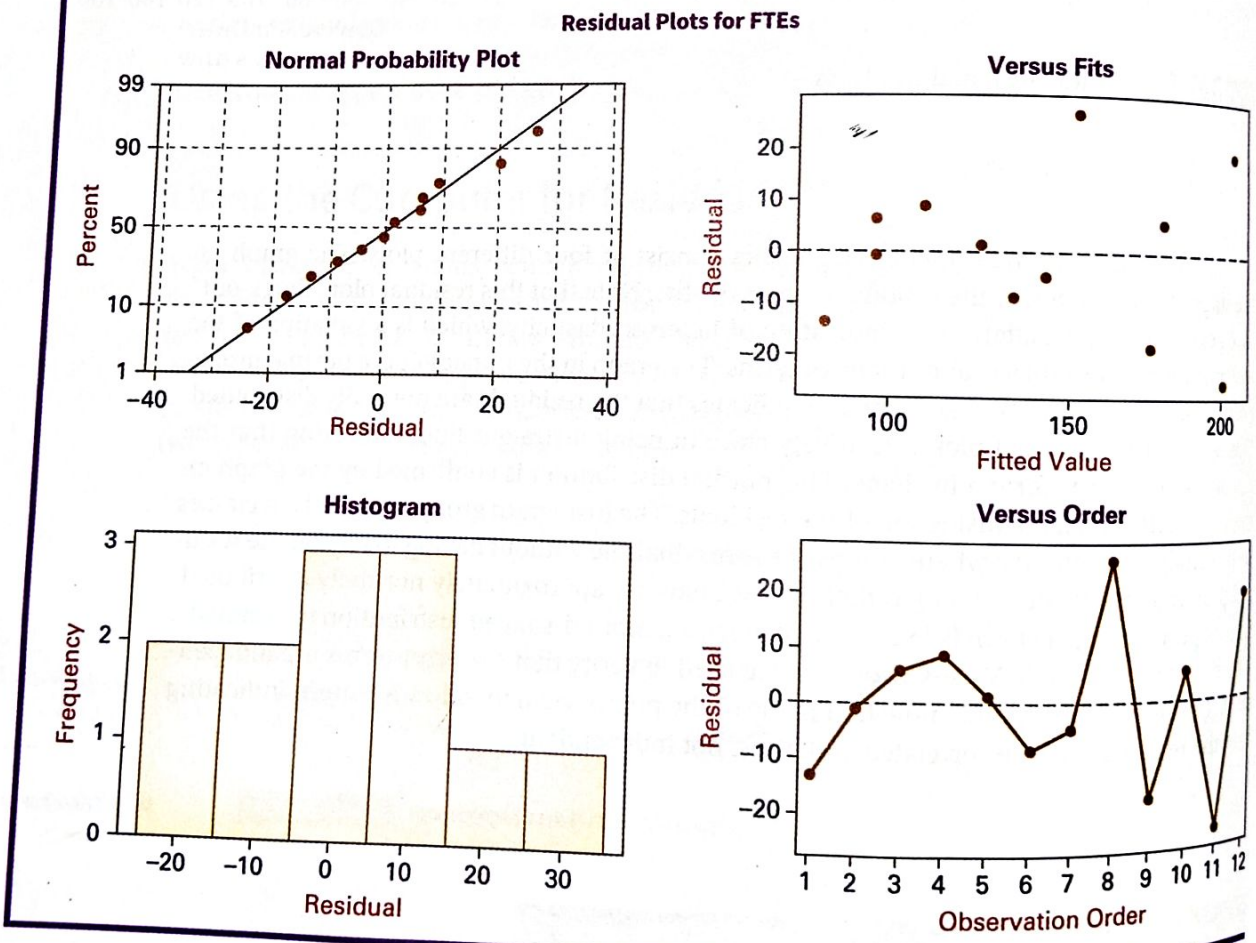
DEMONSTRATION PROBLEM 12.2

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. The Versus Order plot reveals that the residuals appear to be relatively independent.



12.4 Problems

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

x	18	11	22	15	8
y	52	41	61	49	26

12.14. Solve for the predicted values of y and the residuals for the data given below

x	14	23	30	10	22
y	20	18	25	22	27

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	9	46	70	88	112	128

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

Bond Rate (%)	4	11	8	14	6
Prime Interest Rate (%)	15	5	7	3	6

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	4%	11%	8%	14%	6%
Prime Interest Rate	15%	5%	7%	3%	6%

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 (1,000s)	Firm Births (10,000s)
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	1245
2.31	425
2.45	1346
2.52	973
2.19	255
2.55	865
2.40	1080
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.

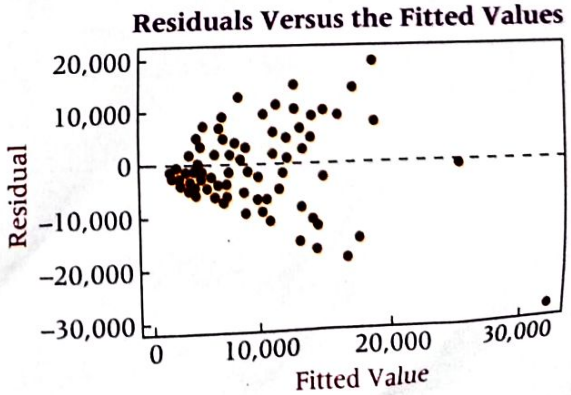
12.21. Graph the following residuals, and indicate which of the assumptions underlying regression appear to be in jeopardy on the basis of the graph.

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 business analyst 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 sometimes easier to compute. The computational formula for SSE follows.

Computational Formula for SSE

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

TABLE 12.6 Determining SSE for the Airline Cost Example

Number of Passengers x	Cost ($1,000s) 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

For the airline cost example,

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

$$b_0 = 1.5697928$$

$$b_1 = .0407016^*$$

$$\Sigma y = 56.69$$

$$\Sigma xy = 4462.22$$

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

$$= 270.9251 - (1.5697928)(56.69) - (.0407016)(4462.22) = .31405$$

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*. The standard error of the estimate for simple regression follows.

Standard Error of the Estimate

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

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

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

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 analyst 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 here 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 8 out of 12 (67%) of the residuals are within $\pm 1s_e$ and 100% of the residuals are within $2s_e$. The standard error of the estimate provides a single measure of error, which, if the analyst has enough background in the area being analyzed, can be used to understand the magnitude of errors in the model. In addition, some analysts use the standard error of the estimate to identify outliers. They do so by looking for data that are outside $\pm 2s_e$ or $\pm 3s_e$.

*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.

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.

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 $\pm 1s_e$ and 100% are within $\pm 2s_e$. 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 $\pm 1s_e$?

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 $\pm 1s_e$ and $\pm 2s_e$. 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?

Annual Sales ($ billions)	Annual Volume (millions of shares)
10.5	728.6
48.1	497.9
64.8	439.1
20.1	377.9
11.4	375.5
123.8	363.8
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 analyst 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 used vehicular weight to predict mission cost searched for regression models that have an r^2 of .90 or higher. However, a business analyst 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

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

Solving for r^2 yields Formula 12.5.

Coefficient of Determination

$$r^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}} \qquad (12.5)$$

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

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 10.1% 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

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

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 business analyst 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 ($1,000s)
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 an analyst 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 analyst 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 business analyst can investigate 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 analyst 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

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

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

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

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

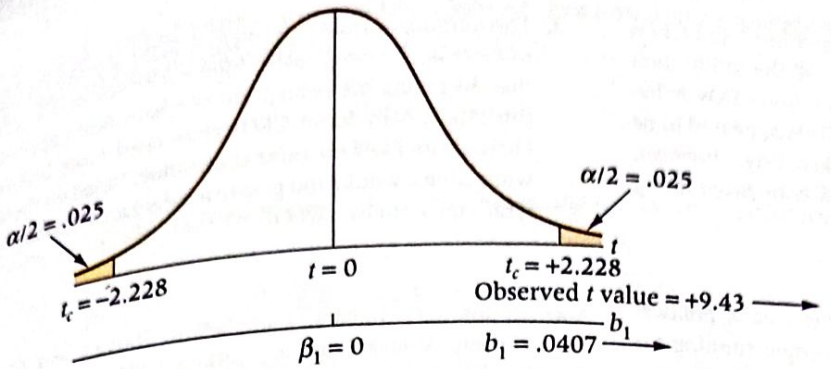


FIGURE 12.14 ***t* Test of Slope from Airline Cost Example**

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

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

As shown in **Figure 12.14**, the t value calculated from the sample slope falls in the rejection region and the p-value is .0000027. 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 the y model does not. Failure to reject the null hypothesis in this test causes the analyst to conclude that the regression model has no predictability of the dependent variable, and the model, therefore, has little or no use.

Thinking Critically About Statistics in Business Today

Are Facial Characteristics Correlated with CEO Traits?

Researchers John R. Graham, Campbell R. Harvey, and Manju Puri, all of the Fuqua School of Business at Duke, conducted a study using almost 2000 participants in an effort to determine if facial characteristics are related to various CEO traits. In one experiment of the study, the researchers showed pictures of 138 CEOs to 230 study participants who were asked to rate each CEO picture in terms of four attributes: competence, attractiveness, trustworthiness, and likeability. The results of the study showed that all four traits are positively correlated. That is, if a CEO (based on the picture) was rated as high on competence, he was also rated high on each of attractiveness, trustworthiness, and likeability. The largest correlation was between trustworthiness and likeability, and the smallest correlation was between trustworthiness and attractiveness. These ratings on each of the four traits were also analyzed to determine if there was a correlation with total sales of the CEO's firm and with CEO income. The results showed that there was a small positive correlation between CEO ratings on competence and company sales. There was also a small positive correlation between CEO ratings on

competence and their income. In another experiment, 138 CEOs were rated on being "baby-faced." Analysis of the study data showed that there was a positive correlation between CEOs' baby-faced rating and likability. That is, the more CEOs appeared to be baby-faced, the higher they were rated in likeability. However, there was a negative correlation between CEOs' baby-faced rating and competence.

Things to Ponder

1. Similar studies have been conducted in the area of political science to determine the electability of people running for office. What do you think is the real impact of studies like this in business?
2. The authors of the study suggest that baby-faced people tend to have large, round eyes, high eyebrows, and a small chin, thereby giving the perception of a baby-faced appearance. In this study, baby-faced CEOs were rated more highly on one attribute and low on another attribute. Based on these results, what advice would you give to a baby-faced business manager who aspires to be a CEO?

Source: John R. Graham, Campbell R. Harvey, and Manju Puri, "A Corporate Beauty Contest," working paper (15906) in the NBER Working Paper Series, National Bureau of Economic Research, at https://www.nber.org/papers/w15906.pdf

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

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

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 analysis of variance (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

$$F = \frac{2.7980/1}{.3141/10} = \frac{2.7980}{.03141} = 89.08$$

The difference between this value (89.08) 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 also unlikely 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 business analyst 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 analyst, 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.