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

Video

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.

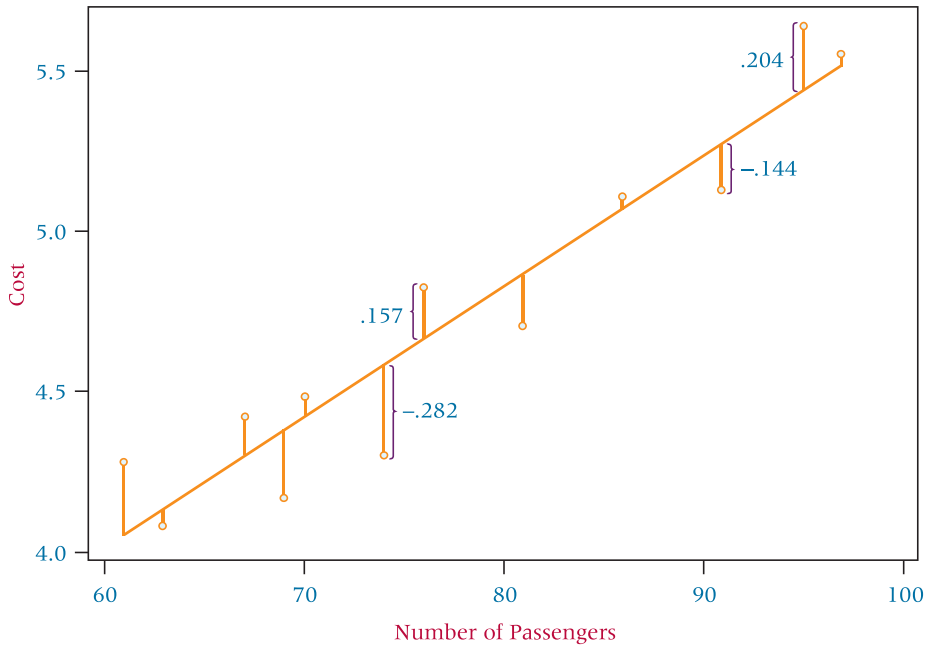
Note that the sum of the residuals is approximately zero. Except for rounding error, the sum of the residuals is *always zero.* The reason is that a residual is geometrically the vertical distance from the regression line to a data point. The equations used to solve for the slope

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

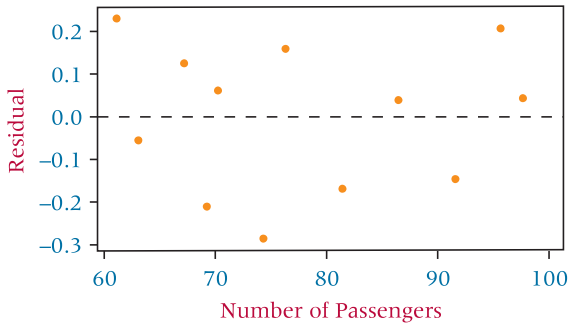


FIGURE 12.9

Nonlinear Residual Plot

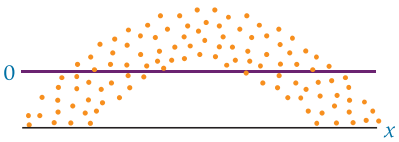
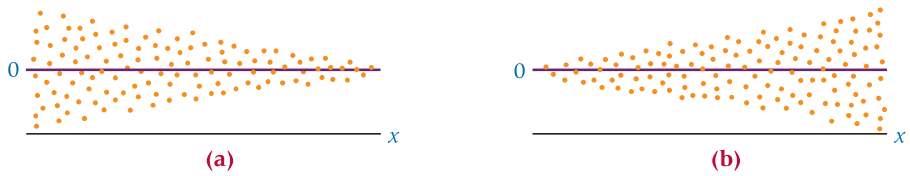


FIGURE 12.10

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

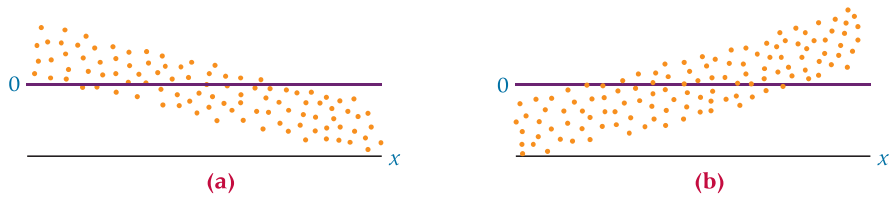
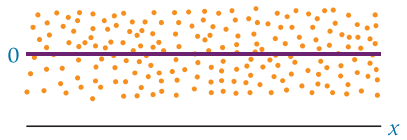


FIGURE 12.12

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