Steeply Sloped Plots

Bare ground (%)	5	5	10	15
Concentration	100	250	300	600
Bare ground (%)	20	25	20	30
Concentration	500	500	900	800
Bare ground (%)	35	40	35	
Concentration	1100	1200	1000	

a. Using the data for steeply sloped plots, find the equation of the least-squares line for predicting y = runoff sediment concentration using x = percentage of bare ground.
b. What would you predict runoff sediment concentration to be for a steeply sloped plot with 18% bare ground?
c. Would you recommend using the least-squares equation from Part (a) to predict runoff sediment concentration for gradually sloped plots? If so, explain why it would be appropriate to do so. If not, provide an alternative way to make such predictions.

5.25 Explain why it can be dangerous to use the least-squares line to obtain predictions for x values that are substantially larger or smaller than those contained in the sample.

5.26 The sales manager of a large company selected a random sample of $n = 10$ salespeople and determined for each one the values of x = years of sales experience and y = annual sales (in thousands of dollars). A scatterplot of the resulting (x, y) pairs showed a linear pattern.

a. Suppose that the sample correlation coefficient is $r = .75$ and that the average annual sales is $\bar{y} = 100$. If a particular salesperson is 2 standard deviations above the mean in terms of experience, what would you predict for that person's annual sales?
b. If a particular person whose sales experience is 1.5 standard deviations below the average experience is predicted to have an annual sales value that is 1 standard deviation below the average annual sales, what is the value of r?

5.27 Explain why the slope b of the least-squares line always has the same sign (positive or negative) as does the sample correlation coefficient r.

5.28 ● The accompanying data resulted from an experiment in which weld diameter x and shear strength y (in pounds) were determined for five different spot welds on steel. A scatterplot shows a strong linear pattern. With $\sum(x - \bar{x})^2 = 1000$ and $\sum(x - \bar{x})(y - \bar{y}) = 8577$, the least-squares line is $\hat{y} = -936.22 + 8.577x$.

x	200.1	210.1	220.1	230.1	240.0
y	813.7	785.3	960.4	1118.0	1076.2

a. Because 1 lb = 0.4536 kg, strength observations can be re-expressed in kilograms through multiplication by this conversion factor: new $y = 0.4536$(old y). What is the equation of the least-squares line when y is expressed in kilograms?
b. More generally, suppose that each y value in a data set consisting of n (x, y) pairs is multiplied by a conversion factor c (which changes the units of measurement for y). What effect does this have on the slope b (i.e., how does the new value of b compare to the value before conversion), on the intercept a, and on the equation of the least-squares line? Verify your conjectures by using the given formulas for b and a. (Hint: Replace y with cy, and see what happens—and remember, this conversion will affect $\bar{y}$.)

Bold exercises answered in back ● Data set available online ✦ Video Solution available

5.3 Assessing the Fit of a Line

Once the least-squares regression line has been obtained, the next step is to examine how effectively the line summarizes the relationship between x and y. Important questions to consider are

1. Is a line an appropriate way to summarize the relationship between the two variables?
2. Are there any unusual aspects of the data set that we need to consider before proceeding to use the regression line to make predictions?
3. If we decide that it is reasonable to use the regression line as a basis for prediction, how accurate can we expect predictions based on the regression line to be?

In this section, we look at graphical and numerical methods that will allow us to answer these questions. Most of these methods are based on the vertical deviations of the data

points from the regression line. These vertical deviations are called *residuals*, and each represents the difference between an actual y value and the corresponding predicted value, $\hat{y}$, that would result from using the regression line to make a prediction.

Predicted Values and Residuals

The predicted value corresponding to the first observation in a data set is obtained by substituting that value, x_1, into the regression equation to obtain $\hat{y}_1$, where

$$\hat{y}_1 = a + bx_1$$

The difference between the actual y value for the first observation, y_1, and the corresponding predicted value is

$$y_1 - \hat{y}_1$$

This difference, called a *residual,* is the vertical deviation of a point in the scatterplot from the regression line.

An observation falling above the line results in a positive residual, whereas a point falling below the line results in a negative residual. This is shown in Figure 5.14.

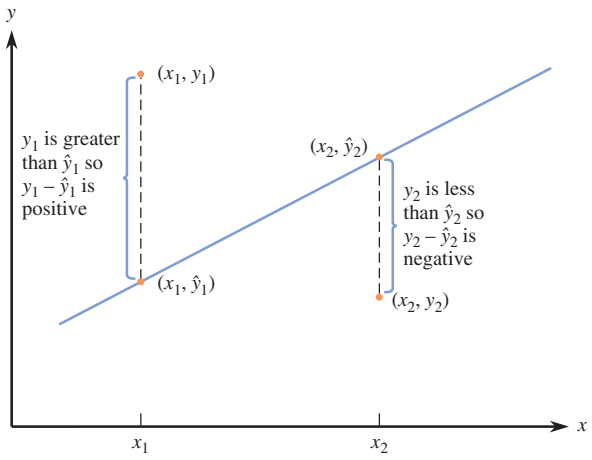


FIGURE 5.14
Positive and negative deviations from the least-squares line (residuals).

DEFINITION

The **predicted** or **fitted values** result from substituting each sample x value in turn into the equation for the least-squares line. This gives

$$\begin{aligned}\hat{y}_1 &= \textit{first predicted value} = a + bx_1\\ \hat{y}_2 &= \textit{second predicted value} = a + bx_2\\ &\vdots\\ \hat{y}_n &= \textit{nth predicted value} = a + bx_n\end{aligned}$$

The **residuals** from the least-squares line are the n quantities

$$\begin{aligned}y_1 - \hat{y}_1 &= \textit{first residual}\\ y_2 - \hat{y}_2 &= \textit{second residual}\\ &\vdots\\ y_n - \hat{y}_n &= \textit{nth residual}\end{aligned}$$

Each residual is the difference between an observed y value and the corresponding predicted y value.

EXAMPLE 5.7 It May Be a Pile of Debris to You, but It Is Home to a Mouse

● The accompanying data is a subset of data read from a scatterplot that appeared in the paper "Small Mammal Responses to fine Woody Debris and Forest Fuel Reduction in Southwest Oregon" (*Journal of Wildlife Management* [2005]: 625–632). The authors of the paper were interested in how the distance a deer mouse will travel for food is related to the distance from the food to the nearest pile of fine woody debris. Distances were measured in meters. The data are given in Table 5.1.

TABLE 5.1 Predicted Values and Residuals for the Data of Example 5.7

Distance From Debris (x)	Distance Traveled (y)	Predicted Distance Traveled ($\hat{y}$)	Residual $(y - \hat{y})$
6.94	0.00	14.76	−14.76
5.23	6.13	9.23	−3.10
5.21	11.29	9.16	2.13
7.10	14.35	15.28	−0.93
8.16	12.03	18.70	−6.67
5.50	22.72	10.10	12.62
9.19	20.11	22.04	−1.93
9.05	26.16	21.58	4.58
9.36	30.65	22.59	8.06

Minitab was used to fit the least-squares regression line. Partial computer output follows:

Regression Analysis: Distance Traveled versus Distance to Debris

The regression equation is

Distance Traveled = −7.7 + 3.23 Distance to Debris

Predictor	Coef	SE Coef	T	P
Constant	−7.69	13.33	−0.58	0.582
Distance to Debris	3.234	1.782	1.82	0.112

S = 8.67071 R-Sq = 32.0% R-Sq(adj) = 22.3%

The resulting least-squares line is $\hat{y} = -7.69 + 3.234x$.

A plot of the data that also includes the regression line is shown in Figure 5.15. The residuals for this data set are the signed vertical distances from the points to the line.

● Data set available online

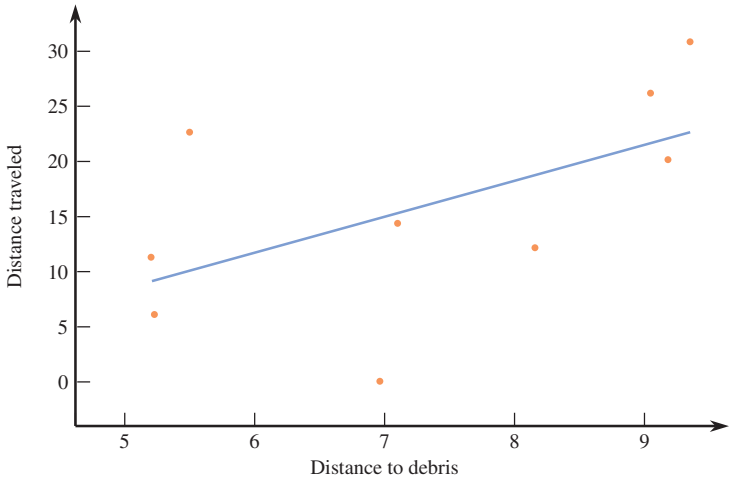


FIGURE 5.15
Scatterplot for the data of Example 5.7.

For the mouse with the smallest x value (the third observation with $x_3 = 5.21$ and $y_3 = 11.29$), the corresponding predicted value and residual are

$$\text{predicted value} = \hat{y}_3 = -7.69 + 3.234(x_3) = -7.69 + 3.234(5.21) = 9.16$$

$$\text{residual} = y_3 - \hat{y}_3 = 11.29 - 9.16 = 2.13$$

The other predicted values and residuals are computed in a similar manner and are included in Table 5.1.

Computing the predicted values and residuals by hand can be tedious, but Minitab and other statistical software packages, as well as many graphing calculators, include them as part of the output, as shown in Figure 5.16. The predicted values and residuals can be found in the table at the bottom of the Minitab output in the columns labeled "Fit" and "Residual," respectively.

The regression equation is
Distance Traveled = – 7.7 + 3.23 Distance to Debris

Predictor	Coef	SE Coef	T	P
Constant	–7.69	13.33	–0.58	0.582
Distance to Debris	3.234	1.782	1.82	0.112

S = 8.67071 R-Sq = 32.0% R-Sq(adj) = 22.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	247.68	247.68	3.29	0.112
Residual Error	7	526.27	75.18		
Total	8	773.95			

Obs	Distance to Debris	Distance Traveled	Fit	SE Fit	Residual	St Resid
1	6.94	0.00	14.76	2.96	–14.76	–1.81
2	5.23	6.13	9.23	4.69	–3.10	–0.42
3	5.21	11.29	9.16	4.72	2.13	0.29
4	7.10	14.35	15.28	2.91	–0.93	–0.11
5	8.16	12.03	18.70	3.27	–6.67	–0.83
6	5.50	22.72	10.10	4.32	12.62	1.68
7	9.19	20.11	22.04	4.43	–1.93	–0.26
8	9.05	26.16	21.58	4.25	4.58	0.61
9	9.36	30.65	22.59	4.67	8.06	1.10

FIGURE 5.16
Minitab output for the data of Example 5.7.

Plotting the Residuals

A careful look at residuals can reveal many potential problems. A *residual plot* is a good place to start when assessing the appropriateness of the regression line.

DEFINITION

A **residual plot** is a scatterplot of the (x, residual) pairs. Isolated points or a pattern of points in the residual plot indicate potential problems.

A desirable residual plot is one that exhibits no particular pattern, such as curvature. Curvature in the residual plot is an indication that the relationship between x and y is not linear and that a curve would be a better choice than a line for describing the relationship between x and y. This is sometimes easier to see in a residual plot than in a scatterplot of y versus x, as illustrated in Example 5.8.

EXAMPLE 5.8 Heights and Weights of American Women

● Consider the accompanying data on x = height (in inches) and y = average weight (in pounds) for American females, age 30–39 (from *The World Almanac and Book of Facts*). The scatterplot displayed in Figure 5.17(a) appears rather straight. However, when the residuals from the least-squares line ($\hat{y} = 98.23 + 3.59x$) are plotted (Figure 5.17(b)), substantial curvature is apparent (even though $r \approx .99$). It is not accurate to say that weight increases in direct proportion to height (linearly with height). Instead, average weight increases somewhat more rapidly for relatively large heights than it does for relatively small heights.

● Data set available online

x	58	59	60	61	62	63	64	65
y	113	115	118	121	124	128	131	134
x	66	67	68	69	70	71	72	
y	137	141	145	150	153	159	164	

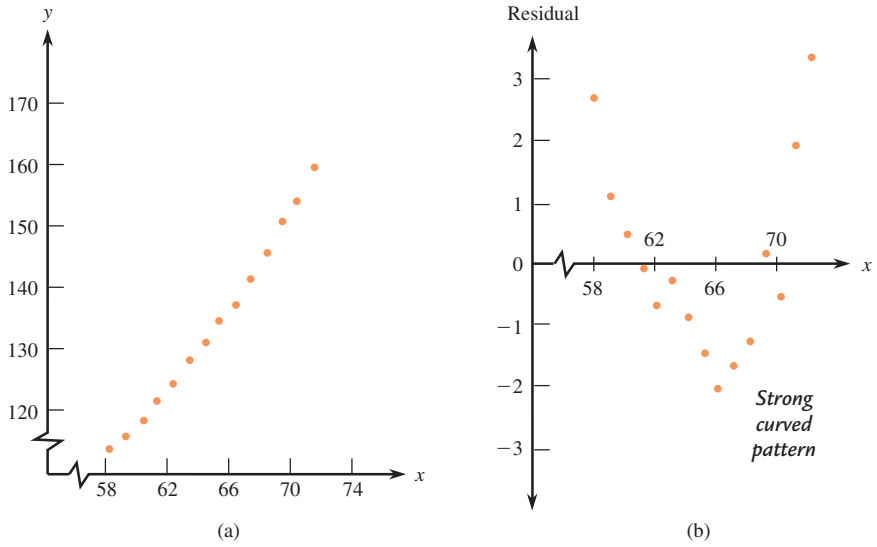


FIGURE 5.17
Plots for the data of Example 5.8: (a) scatterplot; (b) residual plot.

There is another common type of residual plot—one that plots the residuals versus the corresponding $\hat{y}$ values rather than versus the x values. Because $\hat{y} = a + bx$ is simply a linear function of x, the only real difference between the two types of residual plots is the scale on the horizontal axis. The pattern of points in the residual plots will be the same, and it is this pattern of points that is important, not the scale. Thus the two plots give equivalent information, as can be seen in Figure 5.18, which gives both plots for the data of Example 5.7.

It is also important to look for unusual values in the scatterplot or in the residual plot. A point falling far above or below the horizontal line at height 0 corresponds to a large residual, which may indicate some type of unusual behavior, such as a recording error, a nonstandard experimental condition, or an atypical experimental subject. A point whose x value differs greatly from others in the data set may have exerted excessive influence in determining the fitted line. One method for assessing the impact of such an isolated point on the fit is to delete it from the data set, recompute the best-fit line, and evaluate the extent to which the equation of the line has changed.

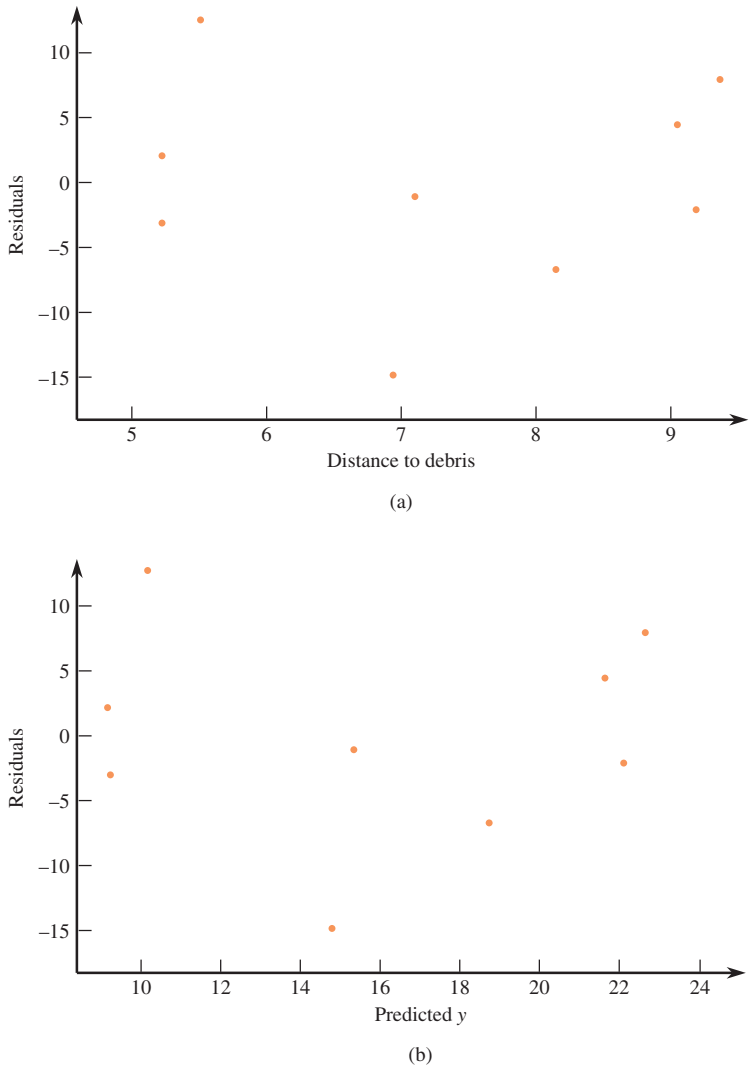


FIGURE 5.18
Plots for the data of Example 5.7.
(a) Plot of residuals versus x;
(b) plot of residuals versus $\hat{y}$.

EXAMPLE 5.9 Older Than Your Average Bear

● The accompanying data on x = age (in years) and y = weight (in kg) for 12 black bears appeared in the paper "Habitat Selection by Black Bears in an Intensively Logged Boreal Forest" (*Canadian Journal of Zoology* [2008]: 1307–1316).

A scatterplot and residual plot are shown in Figures 5.19(a) and 5.19(b), respectively. One bear in the sample was much older than the other bears (bear 3 with an age of $x = 28.5$ years and a weight of $y = 62.00$ kg). This results in a point in the scatterplot that is far to the right of the other points in the scatterplot. Because the least-squares line minimizes the sum of squared residuals, the line is pulled toward this observation. This single observation plays a big role in determining the slope of the least-squares line, and it is therefore called an *influential observation.* Notice that this influential observation is not necessarily one with a large residual, because the least-squares line actually passes near this point. Figure 5.20 shows what happens when the influential observation is removed from the data set. Both the slope and intercept of the least-squares line are quite different from the slope and intercept of the line with this influential observation included.

● Data set available online

Bear	Age	Weight
1	10.5	54
2	6.5	40
3	28.5	62
4	10.5	51
5	6.5	55
6	7.5	56
7	6.5	62
8	5.5	42
9	7.5	40
10	11.5	59
11	9.5	51
12	5.5	50

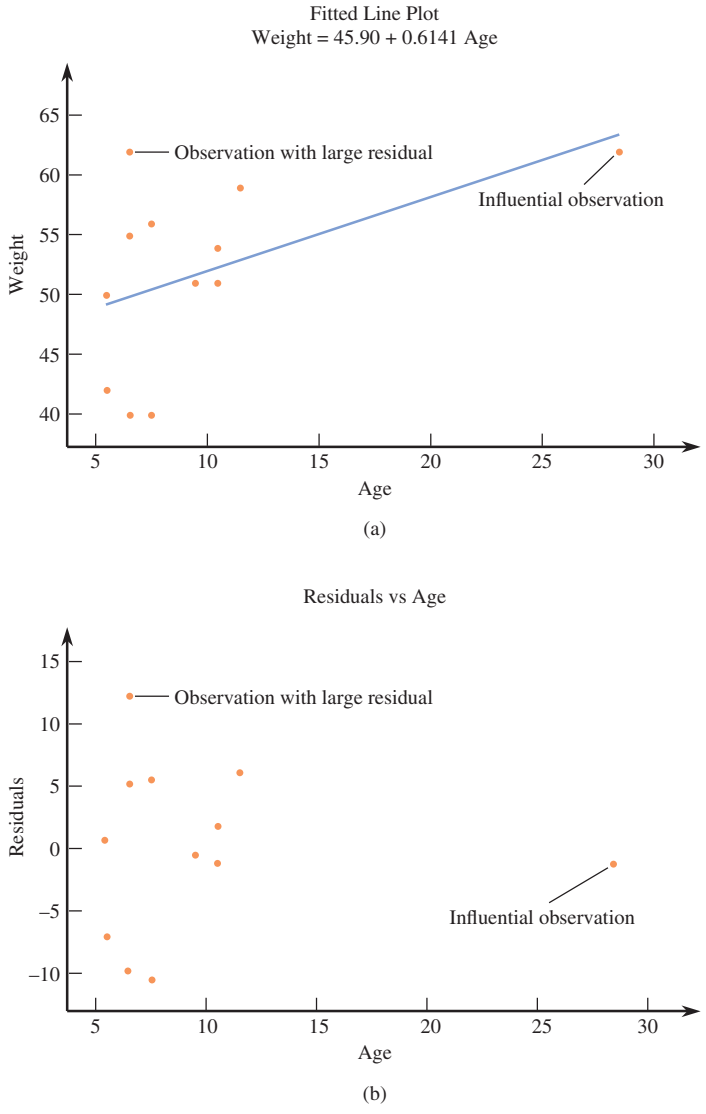


FIGURE 5.19
Minitab plots for the bear data of Example 5.9: (a) scatterplot; (b) residual plot.

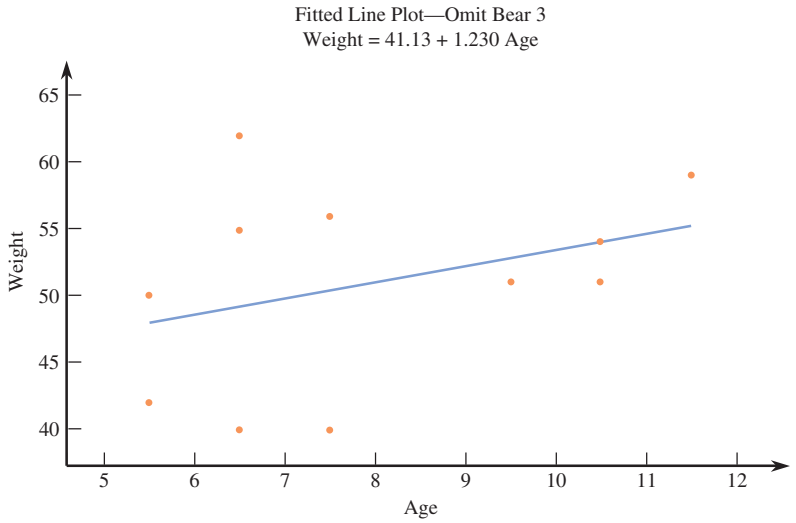


FIGURE 5.20
Scatterplot and least-squares line with bear 3 removed from data set.

Some points in the scatterplot may fall far from the least-squares line in the y direction, resulting in a large residual. These points are sometimes referred to as outliers. In this example, the observation with the largest residual is bear 7 with an age of $x = 6.5$ years and a weight of $y = 62.00$ kg. This observation is labeled in Figure 5.19. Even though this observation has a large residual, this observation is not influential. The equation of the least-squares line for the data set consisting of all 12 observations is $\hat{y} = 45.90 + 0.6141x$, which is not much different from the equation that results from deleting bear 7 from the data set ($\hat{y} = 43.81 + 0.7131x$).

Unusual points in a bivariate data set are those that fall away from most of the other points in the scatterplot in either the x direction or the y direction.

An observation is potentially an **influential observation** if it has an x value that is far away from the rest of the data (separated from the rest of the data in the x direction). To determine if the observation is in fact influential, we assess whether removal of this observation has a large impact on the value of the slope or intercept of the least-squares line.

An observation is an **outlier** if it has a large residual. Outlier observations fall far away from the least-squares line in the y direction.

Careful examination of a scatterplot and a residual plot can help us determine the appropriateness of a line for summarizing a relationship. If we decide that a line is appropriate, the next step is to think about assessing the accuracy of predictions based on the least-squares line and whether these predictions (based on the value of x) are better in general than those made without knowledge of the value of x. Two numerical measures that are helpful in this assessment are the coefficient of determination and the standard deviation about the regression line.

Coefficient of Determination

Suppose that we would like to predict the price of homes in a particular city. A random sample of 20 homes that are for sale is selected, and y = price and x = size (in square feet) are recorded for each house in the sample. There will be variability in house price (the houses will differ with respect to price), and it is this variability that

makes accurate prediction of price a challenge. How much of the variability in house price can be explained by the fact that price is related to house size and that houses differ in size? If differences in size account for a large proportion of the variability in price, a price prediction that takes house size into account is a big improvement over a prediction that is not based on size.

The **coefficient of determination** is a measure of the proportion of variability in the y variable that can be "explained" by a linear relationship between x and y.

DEFINITION

The **coefficient of determination**, denoted by r^2, gives the proportion of variation in y that can be attributed to an approximate linear relationship between x and y.

The value of r^2 is often converted to a percentage (by multiplying by 100) and interpreted as the percentage of variation in y that can be explained by an approximate linear relationship between x and y.

To understand how r^2 is computed, we first consider variation in the y values. Variation in y can effectively be explained by an approximate straight-line relationship when the points in the scatterplot fall close to the least-squares line—that is, when the residuals are small in magnitude. A natural measure of variation about the least-squares line is the sum of the squared residuals. (Squaring before combining prevents negative and positive residuals from counteracting one another.) A second sum of squares assesses the total amount of variation in observed y values by considering how spread out the y values are from the mean y value.

DEFINITION

The **total sum of squares**, denoted by **SSTo**, is defined as

$$\text{SSTo} = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2 = \Sigma(y - \bar{y})^2$$

The **residual sum of squares** (sometimes referred to as the error sum of squares), denoted by **SSResid**, is defined as

$$\text{SSResid} = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2 = \Sigma(y - \hat{y})^2$$

These sums of squares can be found as part of the regression output from most standard statistical packages or can be obtained using the following computational formulas:

$$\text{SSTo} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

$$\text{SSResid} = \Sigma y^2 - a\Sigma y - b\Sigma xy$$

EXAMPLE 5.10 Revisiting the Deer Mice Data

Figure 5.21 displays part of the Minitab output that results from fitting the least-squares line to the data on y = distance traveled for food and x = distance to nearest woody debris pile from Example 5.7. From the output,

SSTo = 773.95 and SSResid = 526.27

Notice that SSResid is fairly large relative to SSTo.

Regression Analysis: Distance Traveled versus Distance to Debris

The regression equation is
Distance Traveled = – 7.7 + 3.23 Distance to Debris

Predictor	Coef	SE Coef	T	P
Constant	–7.69	13.33	–0.58	0.582
Distance to Debris	3.234	1.782	1.82	0.112

S = 8.67071 R-Sq = 32.0% R-Sq(adj) = 22.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	247.68	247.68	3.29	0.112
Residual Error	7	526.27	75.18		
Total	8	773.95			

SSTo SSResid

FIGURE 5.21
Minitab output for the data of Example 5.10.

The residual sum of squares is the sum of squared vertical deviations from the least-squares line. As Figure 5.22 illustrates, SSTo is also a sum of squared vertical deviations from a line—the horizontal line at height $\bar{y}$. The least-squares line is, by definition, the one having the smallest sum of squared deviations. It follows that SSResid $\leq$ SSTo. The two sums of squares are equal only when the least-squares line *is* the horizontal line.

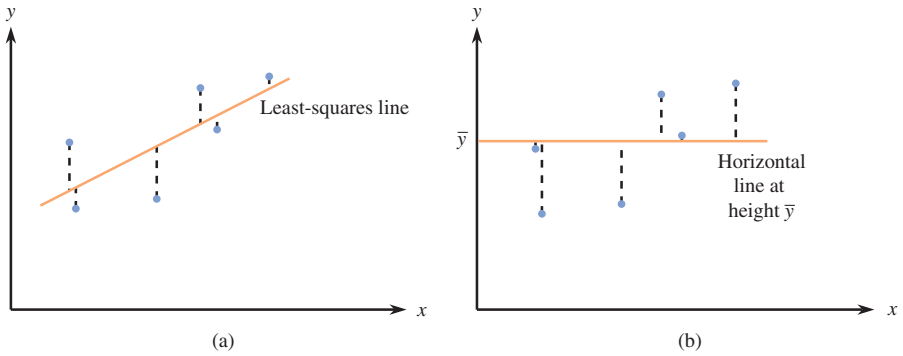


FIGURE 5.22
Interpreting sums of squares: (a) SSResid = sum of squared vertical deviations from the least-squares line; (b) SSTo = sum of squared vertical deviations from the horizontal line at height $\bar{y}$.

SSResid is often referred to as a measure of unexplained variation—the amount of variation in y that cannot be attributed to the linear relationship between x and y. The more the points in the scatterplot deviate from the least-squares line, the larger the value of SSResid and the greater the amount of y variation that cannot be explained by the approximate linear relationship. Similarly, SSTo is interpreted as a measure of total variation. The larger the value of SSTo, the greater the amount of variability in $y_1, y_2, \ldots, y_n$.

The ratio SSResid/SSTo is the fraction or proportion of total variation that is unexplained by a straight-line relation. Subtracting this ratio from 1 gives the proportion of total variation that *is* explained:

The coefficient of determination is computed as

$$r^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}}$$

Multiplying r^2 by 100 gives the percentage of y variation attributable to the approximate linear relationship. The closer this percentage is to 100%, the more successful is the relationship in explaining variation in y.

EXAMPLE 5.11 r^2 for the Deer Mice Data

For the data on distance traveled for food and distance to nearest debris pile from Example 5.10, we found SSTo = 773.95 and SSResid = 526.27. Thus

$$r^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}} = 1 - \frac{526.27}{773.95} = .32$$

This means that only 32% of the observed variability in distance traveled for food can be explained by an approximate linear relationship between distance traveled for food and distance to nearest debris pile. Note that the r^2 value can be found in the Minitab output of Figure 5.21, labeled "R-Sq."

The symbol r was used in Section 5.1 to denote Pearson's sample correlation coefficient. It is not coincidental that r^2 is used to represent the coefficient of determination. The notation suggests how these two quantities are related:

(correlation coefficient)^2 = coefficient of determination

Thus, if $r = .8$ or $r = -.8$, then $r^2 = .64$, so 64% of the observed variation in the dependent variable can be explained by the linear relationship. Because the value of r does not depend on which variable is labeled x, the same is true of r^2. The coefficient of determination is one of the few quantities computed in a regression analysis whose value remains the same when the roles of dependent and independent variables are interchanged. When $r = .5$, we get $r^2 = .25$, so only 25% of the observed variation is explained by a linear relation. This is why a value of r between $-.5$ and .5 is not considered evidence of a strong linear relationship.

EXAMPLE 5.12 Lead Exposure and Brain Volume

The authors of the paper "Decreased Brain Volume in Adults with Childhood Lead Exposure" (*Public Library of Science Medicine* [May 27, 2008]: e112) studied the relationship between childhood environmental lead exposure and a measure of brain volume change in a particular region of the brain. Data on x = mean childhood blood lead level (μg/dL) and y = brain volume change (percent) read from a graph that appeared in the paper was used to produce the scatterplot in Figure 5.23. The least-squares line is also shown on the scatterplot.

Figure 5.24 displays part of the Minitab output that results from fitting the least-squares line to the data. Notice that although there is a slight tendency for smaller y values (corresponding to a brain volume decrease) to be paired with higher values of mean blood lead levels, the relationship is weak. The points in the plot are widely scattered around the least-squares line.

From the computer output, we see that $100r^2 = 13.6\%$, so $r^2 = .136$. This means that differences in childhood mean blood lead level explain only 13.6% of the variability in adult brain volume change. Because the coefficient of determination is the square of the correlation coefficient, we can compute the value of the correlation

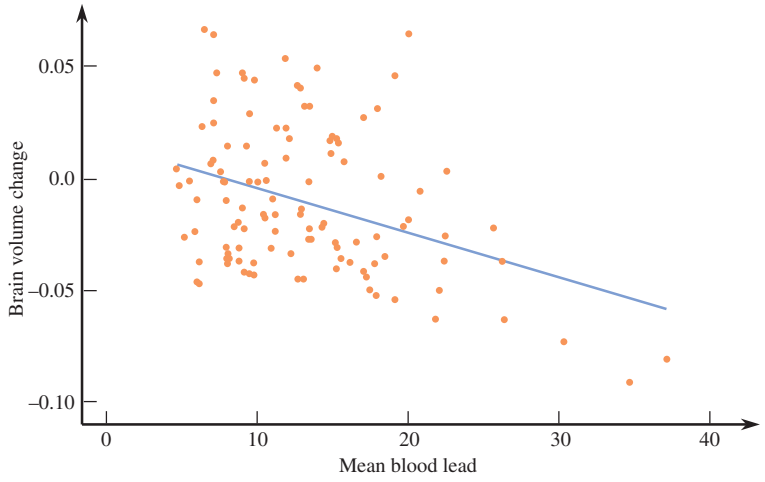


FIGURE 5.23
Scatterplot and least-squares line for the data of Example 5.12.

Regression Analysis: Brain Volume Change versus Mean Blood Lead

The regression equation is
Brain Volume Change = 0.01559 – 0.001993 Mean Blood Lead

S = 0.0310931 R-Sq = 13.6% R-Sq(adj) = 12.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.016941	0.0169410	17.52	0.000
Error	111	0.107313	0.0009668		
Total	112	0.124254			

FIGURE 5.24
Minitab output for the data of Example 5.12.

coefficient by taking the square root of r^2. In this case, we know that the correlation coefficient will be negative (because there is a negative relationship between x and y), so we want the negative square root:

$$r = -\sqrt{.136} = -.369$$

Based on the values of the correlation coefficient and the coefficient of determination, we would conclude that there is a weak negative linear relationship and that childhood mean blood lead level explains only about 13.6% of adult change in brain volume.

Standard Deviation About the Least-Squares Line

The coefficient of determination measures the extent of variation about the best-fit line *relative* to overall variation in y. A high value of r^2 does not by itself promise that the deviations from the line are small in an absolute sense. A typical observation could deviate from the line by quite a bit, yet these deviations might still be small relative to overall y variation.

Recall that in Chapter 4 the sample standard deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

was used as a measure of variability in a single sample; roughly speaking, s is the typical amount by which a sample observation deviates from the mean. There is an analogous measure of variability when a least-squares line is fit.

DEFINITION

The **standard deviation about the least-squares line** is given by

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

Roughly speaking, s_e is the typical amount by which an observation deviates from the least-squares line. Justification for division by $(n-2)$ and the use of the subscript e is given in Chapter 13.

EXAMPLE 5.13 Predicting Graduation Rates

● Consider the accompanying data from 2007 on six-year graduation rate (%), student-related expenditure per full-time student, and median SAT score for the 38 primarily undergraduate public universities and colleges in the United States with enrollments between 10,000 and 20,000 (Source: College Results Online, The Education Trust).

Graduation Rate	Expenditure	Median SAT
81.2	7462	1160
66.8	7310	1115
66.4	6959	1070
66.1	8810	1205
64.9	7657	1135
63.7	8063	1060
62.6	8352	1130
62.5	7789	1200
61.2	8106	1015
59.8	7776	1100
56.6	8515	990
54.8	7037	1085
52.7	8715	1040
52.4	7780	1040
52.4	7198	1105
50.5	7429	975
49.9	7551	1030
48.9	8112	1030
48.1	8149	950
46.5	6744	1010
45.3	8842	1223
45.2	7743	990
43.7	5587	1010
43.5	7166	1010
42.9	5749	950
42.1	6268	955
42.0	8477	985
38.9	7076	990
38.8	8153	990
38.3	7342	910
35.9	8444	1075

● Data set available online

Graduation Rate	Expenditure	Median SAT
32.8	7245	885
32.6	6408	1060
32.3	4981	990
31.8	7333	970
31.3	7984	905
31.0	5811	1010
26.0	7410	1005

Figure 5.25 displays scatterplots of graduation rate versus student-related expenditure and graduation rate versus median SAT score. The least-squares lines and the values of r^2 and s_e are also shown.

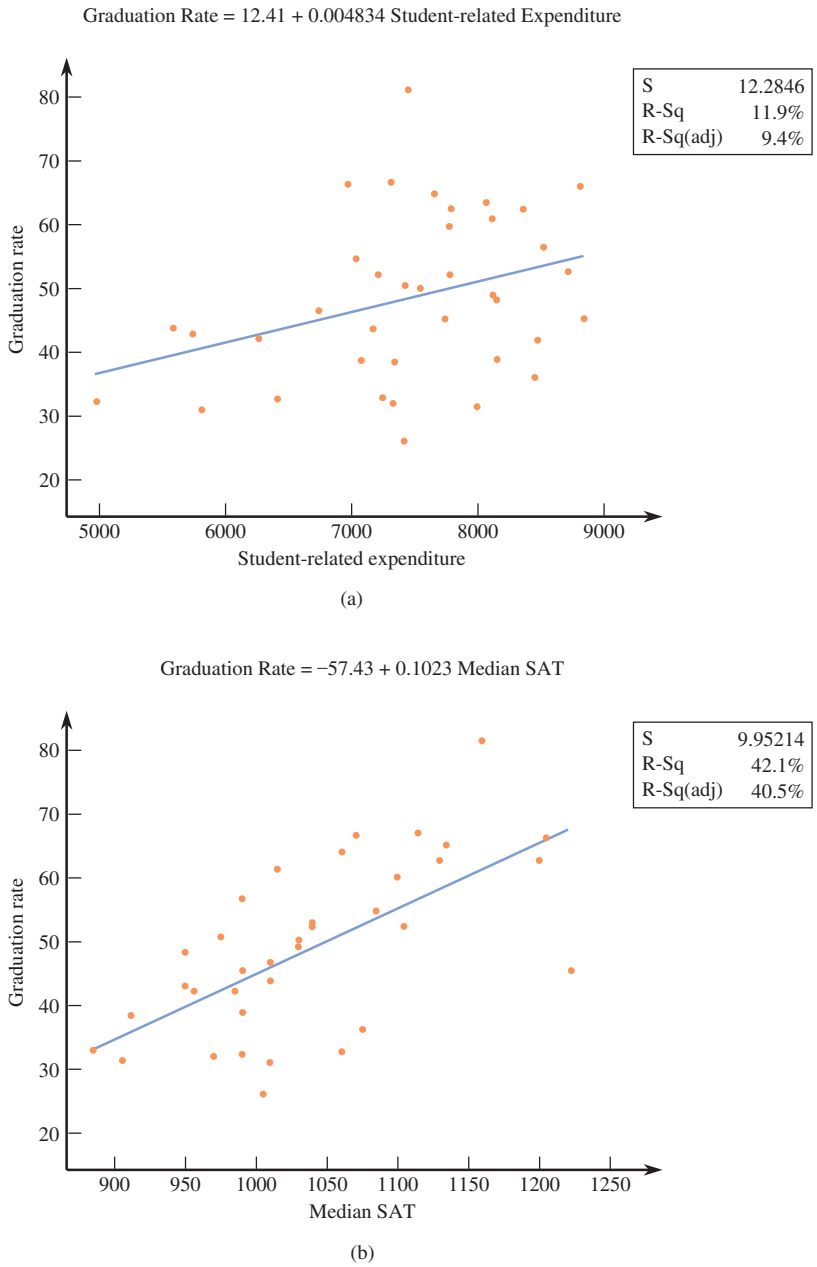


FIGURE 5.25
Scatterplots for the data of Example 5.13: (a) graduation rate versus student-related expenditure; (b) graduation rate versus median SAT.

Notice that while there is a positive linear relationship between student-related expenditure and graduation rate, the relationship is weak. The value of r^2 is only .119 (11.9%), indicating that only about 11.9% of the variability in graduation rate from university to university can be explained by student-related expenditures. The standard deviation about the regression line is $s_e = 12.2846$, which is larger than s_e for the predictor median SAT, a reflection of the fact that the points in the scatterplot of graduation rate versus student-related expenditure tend to fall farther from the regression line than is the case for the line that describes graduation rate versus median SAT. The value of r^2 for graduation rate versus median SAT is .421 (42.1%) and $s_e = 9.95214$, indicating that the predictor median SAT does a better job of explaining variability in graduation rates and the corresponding least-squares line would be expected to produce more accurate estimates of graduation rates than would be the case for the predictor student-related expenditure.

Based on the values of r^2 and s_e, median SAT would be a better choice for predicting graduation rates than student-related expenditures. It is also possible to develop a prediction equation that would incorporate both potential predictors—techniques for doing this are introduced in Chapter 14.

EXERCISES 5.29 - 5.43

5.29 ● The data in the accompanying table is from the paper "Six-Minute Walk Test in Children and Adolescents" (*The Journal of Pediatrics* [2007]: 395–399). Two hundred and eighty boys completed a test that measures the distance that the subject can walk on a flat, hard surface in 6 minutes. For each age group shown in the table, the median distance walked by the boys in that age group is also given.

Age Group	Representative Age (Midpoint of Age Group)	Median Six-minute Walk Distance (meters)
3–5	4	544.3
6–8	7	584.0
9–11	10	667.3
12–15	13.5	701.1
16–18	17	727.6

a. With x = representative age and y = median distance walked in 6 minutes, construct a scatterplot. Does the pattern in the scatterplot look linear?

b. Find the equation of the least-squares regression line that describes the relationship between median distance walked in 6 minutes and representative age.

c. Compute the five residuals and construct a residual plot. Are there any unusual features in the plot?

5.30 ● The paper referenced in the previous exercise also gave the 6-minute walk distances for 248 girls age 3 to 18 years. The median 6-minute walk times for girls for the five age groups were

492.4 578.3 655.8 657.6 660.9

a. With x = representative age and y = median distance walked in 6 minutes, construct a scatterplot. How does the pattern in the scatterplot for girls differ from the pattern in the scatterplot for boys from Exercise 5.29?

b. Find the equation of the least-squares regression line that describes the relationship between median distance walked in 6 minutes and representative age for girls.

c. Compute the five residuals and construct a residual plot. The authors of the paper decided to use a curve rather than a straight line to describe the relationship between median distance walked in 6 minutes and age for girls. What aspect of the residual plot supports this decision?

5.31 ● Data on pollution and cost of medical care for elderly people were given in Exercise 5.14 and are also shown here. The accompanying data are a measure of pollution (micrograms of particulate matter per cubic meter of air) and the cost of medical care per person over age 65 for six geographic regions of the United States.

Region	Pollution	Cost of Medical Care
North	30.0	915
Upper South	31.8	891
Deep South	32.1	968
West South	26.8	972
Big Sky	30.4	952
West	40.0	899

The equation of the least-squares regression line for this data set is $\hat{y} = 1082.2 - 4.691x$, where y = medical cost and x = pollution.

a. Compute the six residuals.

b. What is the value of the correlation coefficient for this data set? Does the value of r indicate that the linear relationship between pollution and medical cost is strong, moderate, or weak? Explain.

c. Construct a residual plot. Are there any unusual features of the plot?

d. The observation for the West, (40.0, 899), has an x value that is far removed from the other x values in the sample. Is this observation influential in determining the values of the slope and/or intercept of the least-squares line? Justify your answer.

5.32 ● Northern flying squirrels eat lichen and fungi, which makes for a relatively low quality diet. The authors of the paper "Nutritional Value and Diet Preference of Arboreal Lichens and Hypogeous Fungi for Small Mammals in the Rocky Mountain" (*Canadian Journal of Zoology* [2008]: 851–862) measured nitrogen intake and nitrogen retention in six flying squirrels that were fed the fungus *Rhizopogon*. Data read from a graph that appeared in the paper are given in the table below. (The negative value for nitrogen retention for the first squirrel represents a net loss in nitrogen.)

Nitrogen Intake, x (grams)	Nitrogen Retention, y (grams)
0.03	−0.04
0.10	0.00
0.07	0.01
0.06	0.01
0.07	0.04
0.25	0.11

a. Construct a scatterplot of these data.

b. Find the equation of the least-squares regression line. Based on this line, what would you predict nitrogen retention to be for a flying squirrel whose nitrogen intake is 0.06 grams? What is the residual associated with the observation (0.06, 0.01)?

c. Look again at the scatterplot from Part (a). Which observation is potentially influential? Explain the reason for your choice.

d. When the potentially influential observation is deleted from the data set, the equation of the least-squares regression line fit to the remaining five observations is $\hat{y} = -0.037 + 0.627x$. Use this equation to predict nitrogen retention for a flying squirrel whose nitrogen intake is 0.06. Is this prediction much different than the prediction made in Part (b)?

5.33 ● The relationship between x = total number of salmon in a creek and y = percentage of salmon killed by bears that were transported away from the stream prior to the bear eating the salmon was examined in the paper "Transportation of Pacific Salmon Carcasses from Streams to Riparian Forests by Bears" (*Canadian Journal of Zoology* [2009]: 195–203). Data for the 10 years from 1999 to 2008 is given in the accompanying table.

Total Number	Percentage Transported
19,504	77.8
3,460	28.7
1,976	28.9
8,439	27.9
11,142	55.3
3,467	20.4
3,928	46.8
20,440	76.3
7,850	40.3
4,134	24.1

a. Construct a scatterplot of the data. Does there appear to be a relationship between the total number of salmon in the stream and the percentage of salmon killed by bears that are transported away from the stream?

b. Find the equation of the least-squares regression line. Draw the regression line for the scatterplot from Part (a).

c. The residuals from the least-squares line are shown in the accompanying table. The observation (3928, 46.8) has a large residual. Is this data point also an influential observation?

Total Number	Percent Transported	Residual
19,504	77.8	3.43
3,460	28.7	0.30
1,976	28.9	4.76
8,439	27.9	−14.76
11,142	55.3	4.89
3,467	20.4	−8.02
3,928	46.8	17.06
20,440	76.3	−0.75
7,850	40.3	−0.68
4,134	24.1	−6.23

d. The two points with unusually large x values (19,504 and 20,440) were not thought to be influential observations even though they are far removed in the x direction from the rest of the points in the scatterplot. Explain why these two points are not influential.

e. Partial Minitab output resulting from fitting the least-squares line is shown here. What is the value of s_e? Write a sentence interpreting this value.

Regression Analysis: Percent Transported versus Total Number

The regression equation is

Percent Transported = 18.5 + 0.00287 Total Number

Predictor	Coef	SE Coef	T	P
Constant	18.483	4.813	3.84	0.005
Total Number	0.0028655	0.0004557	6.29	0.000

S = 9.16217 R-Sq = 83.2% R-Sq(adj) = 81.1%

f. What is the value of r^2 for this data set (see Minitab output in Part (e))? Is the value of r^2 large or small? Write a sentence interpreting the value of r^2.

5.34 ● The paper "Effects of Age and Gender on Physical Performance" (*Age* [2007]: 77–85) describes a study of the relationship between age and 1-hour swimming performance. Data on age and swim distance for over 10,000 men participating in a national long-distance 1-hour swimming competition are summarized in the accompanying table.

Age Group	Representative Age (Midpoint of Age Group)	Average Swim Distance (meters)
20–29	25	3913.5
30–39	35	3728.8
40–49	45	3579.4
50–59	55	3361.9
60–69	65	3000.1
70–79	75	2649.0
80–89	85	2118.4

(continued)

a. Find the equation of the least-squares line with x = representative age and y = average swim distance.

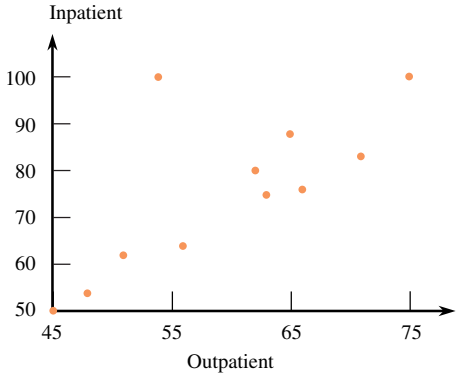
b. Compute the seven residuals and use them to construct a residual plot. What does the residual plot suggest about the appropriateness of using a line to describe the relationship between representative age and swim distance?

c. Would it be reasonable to use the least-squares line from Part (a) to predict the average swim distance for women age 40 to 49 by substituting the representative age of 45 into the equation of the least-squares line? Explain.

5.35 Data on x = representative age and y = 6-minute walk time for boys were given in Exercise 5.29. Compute the values of s_e and r^2 for these data. What do these values tell you about the fit of the least-squares line?

5.36 ● Cost-to-charge ratio (the percentage of the amount billed that represents the actual cost) for inpatient and outpatient services at 11 Oregon hospitals is shown in the following table (Oregon Department of Health Services, 2002). A scatterplot of the data is also shown.

	Cost-to-Charge Ratio	
Hospital	Outpatient Care	Inpatient Care
1	62	80
2	66	76
3	63	75
4	51	62
5	75	100
6	65	88
7	56	64
8	45	50
9	48	54
10	71	83
11	54	100



The least-squares regression line with y = inpatient cost-to-charge ratio and x = outpatient cost-to-charge ratio is $\hat{y} = -1.1 + 1.29x$.

a. Is the observation for Hospital 11 an influential observation? Justify your answer.
b. Is the observation for Hospital 11 an outlier? Explain.
c. Is the observation for Hospital 5 an influential observation? Justify your answer.
d. Is the observation for Hospital 5 an outlier? Explain.

5.37 The article "Examined Life: What Stanley H. Kaplan Taught Us About the SAT" (*The New Yorker* [December 17, 2001]: 86–92) included a summary of findings regarding the use of SAT I scores, SAT II scores, and high school grade point average (GPA) to predict first-year college GPA. The article states that "among these, SAT II scores are the best predictor, explaining 16 percent of the variance in first-year college grades. GPA was second at 15.4 percent, and SAT I was last at 13.3 percent."

a. If the data from this study were used to fit a least-squares line with y = first-year college GPA and x = high school GPA, what would be the value of r^2?

b. The article stated that SAT II was the best predictor of first-year college grades. Do you think that predictions based on a least-squares line with y = first-year college GPA and x = SAT II score would be very accurate? Explain why or why not.

5.38 ● The paper "Accelerated Telomere Shortening in Response to Life Stress" (*Proceedings of the National Academy of Sciences* [2004]: 17312–17315) described a study that examined whether stress accelerates aging at a cellular level. The accompanying data on a measure of perceived stress (x) and telomere length (y) were read from a scatterplot that appeared in the paper. Telomere length is a measure of cell longevity.

Perceived Stress	Telomere Length	Perceived Stress	Telomere Length
5	1.25	20	1.22
6	1.32	20	1.30
6	1.5	20	1.32
7	1.35	21	1.24
10	1.3	21	1.26
11	1	21	1.30
12	1.18	22	1.18
13	1.1	22	1.22
14	1.08	22	1.24
14	1.3	23	1.18
15	0.92	24	1.12
15	1.22	24	1.50
15	1.24	25	0.94
17	1.12	26	0.84
17	1.32	27	1.02
17	1.4	27	1.12
18	1.12	28	1.22
18	1.46	29	1.30
19	0.84	33	0.94

a. Compute the equation of the least-squares line.
b. What is the value of r^2?
c. Does the linear relationship between perceived stress and telomere length account for a large or small proportion of the variability in telomere length? Justify your answer.

5.39 ● The article "California State Parks Closure List Due Soon" (*The Sacramento Bee*, August 30, 2009) gave the following data on y = number of employees in fiscal year 2007–2008 and x = total size of parks (in acres) for the 20 state park districts in California:

Number of Employees, y	Total Park Size, x
95	39,334
95	324
102	17,315
69	8,244
67	620,231
77	43,501
81	8,625
116	31,572
51	14,276
36	21,094
96	103,289
71	130,023
76	16,068
112	3,286
43	24,089

Continued

Number of Employees, y	Total Park Size, x
87	6,309
131	14,502
138	62,595
80	23,666
52	35,833

a. Construct a scatterplot of the data.
b. Find the equation of the least-squares line. Do you think the least-squares line gives accurate predictions? Explain.
c. Delete the observation with the largest x value from the data set and recalculate the equation of the least-squares line. Does this observation greatly affect the equation of the line?

5.40 ● The article referenced in the previous exercise also gave data on the percentage of operating costs covered by park revenues for the 2007–2008 fiscal year.

Number of Employees, x	Percent of Operating Cost Covered by Park Revenues, y
95	37
95	19
102	32
69	80
67	17
77	34
81	36
116	32
51	38
36	40

(continued)

Number of Employees, x	Percent of Operating Cost Covered by Park Revenues, y
96	53
71	31
76	35
112	108
43	34
87	97
131	62
138	36
80	36
52	34

a. Find the equation of the least-squares line relating y = percent of operating costs covered by park revenues and x = number of employees.
b. Based on the values of r^2 and s_e, do you think that the least-squares regression line does a good job of describing the relationship between y = percent of operating costs covered by park revenues and x = number of employees? Explain.
c. The graph in Figure EX5.40 is a scatterplot of y = percent of operating costs covered by park revenues and x = number of employees. The least-squares line is also shown. Which observations are outliers? Do the observations with the largest residuals correspond to the park districts with the largest number of employees?

5.41 A study was carried out to investigate the relationship between the hardness of molded plastic (y, in Brinell units) and the amount of time elapsed since termination of the molding process (x, in hours). Summary quantities include $n = 15$, SSResid = 1235.470, and

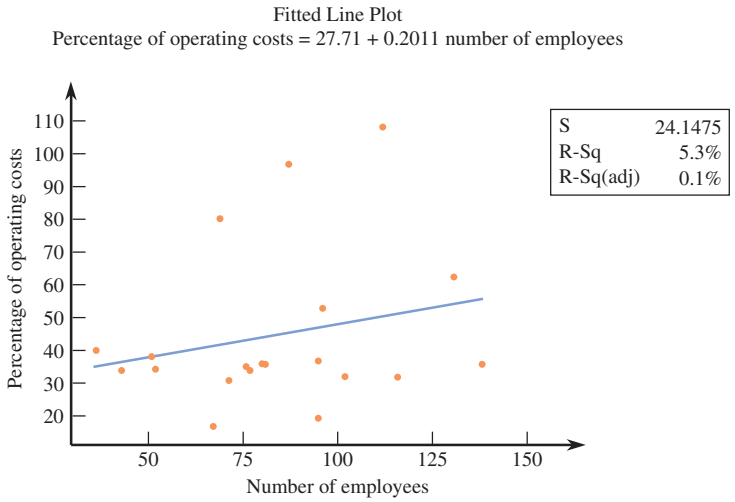


FIGURE EX5.40