Example 1

Estimate a Person's Strength

Picture the Scenario

How can you measure a person's strength? One way is to find the *maximum* number of pounds that the individual can bench press. However, this technique can be risky for people who are unfamiliar with proper lifting techniques or who are inexperienced in using a bench press. Is there a variable that is easier to measure yet is a good predictor of the maximum bench press?

There have been many studies about strength using males but relatively few using females. One exception was a recent study of 57 female athletes in a Georgia high school. Several variables were measured, including ones that are easier and safer to assess than maximum bench press but are thought to correlate highly with it. One such variable is the number of times that a girl can lift a bench press set at only 60 pounds (a relatively low weight) before she becomes too fatigued to lift it again. The data are in the High School Female Athletes data file on the text CD. Let x = number of 60-pound bench presses performed (before fatigue) and let y = maximum bench press.

Questions to Explore

- How well can we predict an athlete's maximum bench press from knowing the number of 60-pound bench presses that she can perform?
- What can we say about the association between these variables in the population?

Thinking Ahead

The bench press variables are quantitative. This chapter presents methods for analyzing the association between two quantitative variables. Like the methods presented in the previous chapter for categorical variables, these methods enable us to answer questions such as:

- Could the variables x and y realistically be independent (in the population), or can we conclude that there is an association between them?
- If the variables are associated, how strong is the association?
- How does the outcome for the response variable depend on the value of the explanatory variable, and which observations are unusual?

The analyses that address these questions are collectively called a **regression analysis**. We'll conduct regression analyses of the data from the female athlete strength study using different variables in several examples in this chapter (Examples 2, 3, 6, 9, 11, 12, 15, and 16).

Example 2

The Strength Study

Picture the Scenario

Another variable in the strength study discussed in Example 1 measured the maximum bench press by giving a girl a 60-pound weight to lift and then increasing the weight in 5-pound increments until the girl could no longer lift it. The response outcome is the maximum weight that the girl bench pressed. Let's look at the High School Female Athletes data file on the text CD. The first four entries of the 57 lines for x = the number of 60-pound bench presses and y = maximum bench press show the values:

x	y
10	80
12	85
20	85
5	65

For the 57 girls in this study, these variables are summarized by

x: mean = 11.0, standard deviation = 7.1

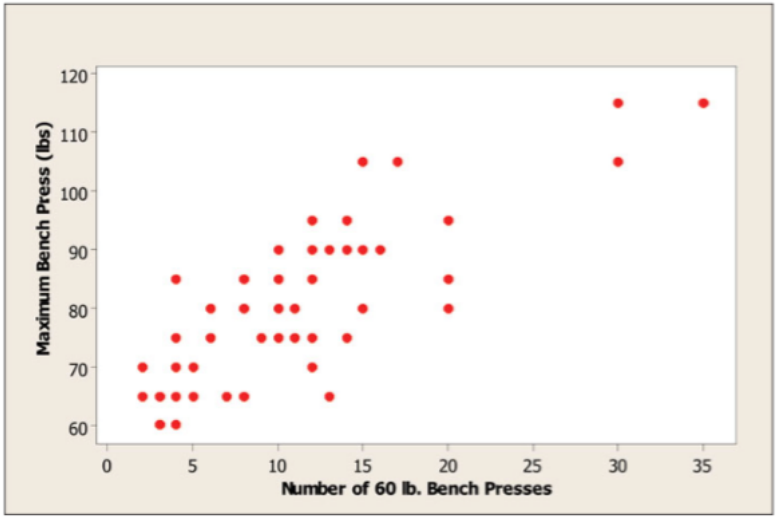
y: mean = 79.9 lbs., standard deviation = 13.3 lbs.

Question to Explore

Using the entire data file for the 57 athletes, construct a scatterplot, and interpret it.

Think It Through

With any statistical software, you can construct a scatterplot after identifying in the data file the variables that play the role of x and y. Figure 12.1 shows the scatterplot that MINITAB produces. It shows that female athletes with higher numbers of 60-pound bench presses also tended to have higher values for the maximum bench press. The data points follow roughly an increasing linear trend. The margin shows a screen shot of the scatterplot from the TI-83+/84.



▲ **Figure 12.1** Scatterplot for y = Maximum Bench Press and x = Number of 60-Pound Bench Presses. **Question** How can you tell from this plot that (a) two athletes had y = maximum bench press of only 60 pounds, (b) 5-pound increments were used in determining maximum bench press—60, 65, 70, and so on?

Insight

The three data points at the upper right represent subjects who could do a large number of 60-pound bench presses. Their maximum bench presses were also high.

Try Exercise 12.9, part a

Example 3

Regression Line Predicting Maximum Bench Press

Picture the Scenario

Let's continue our analysis of the High School Female Athletes data.

Questions to Explore

a. Using software, find the regression line for y = maximum bench press and x = number of 60-pound bench presses.

b. Interpret the slope by comparing the predicted maximum bench press for subjects at the highest and lowest levels of x in the sample (35 and 2).

Think It Through

a. Let's denote the maximum bench press variable by BP and the number of 60-pound bench presses by BP_60. Using software, we pick the regression option with BP as the response variable and BP_60 as the explanatory variable. Table 12.1 shows some output using MINITAB. We'll interpret some of this, such as standard errors and t statistics, later in the chapter. The margin shows TI-83+/84 output.

Table 12.1 MINITAB Printout for Regression Analysis of y = Maximum Bench Press (BP) and x = Number of 60-Pound Bench Presses (BP_60)

```
The regression equation is BP = 63.5 + 1.49 BP_60

Predictor     Coef  SE Coef     T      P

Constant    63.537    1.956 32.48  0.000

BP_60       1.4911   0.1497  9.96  0.000
```

The output tells us that $\hat{y}$ = predicted maximum bench press (BP) relates to x = number of 60-pound bench presses (BP_60) by

$$\texttt{BP} = 63.5 + 1.49\,\texttt{BP\_60}, \text{ that is, } \hat{y} = 63.5 + 1.49x.$$

The y-intercept is 63.5 and the slope is 1.49. These are also shown in the column labeled "Coef," an abbreviation for *coefficient*. The slope appears opposite the variable name for which it is the coefficient, "BP_60." The predictor "Constant" refers to the y-intercept. The margin figure (Figure 12.1 reproduced) plots the regression line through the scatterplot.

b. The slope of 1.49 tells us that the predicted maximum bench press $\hat{y}$ increases by an average of 1 1/2 pounds for every additional 60-pound bench press an athlete can do. The impact on $\hat{y}$ of a 33-unit change in x, from the sample minimum of $x = 2$ to the maximum of $x = 35$, is $33(1.49) = 49.2$ pounds. An athlete who can do thirty-five 60-pound bench presses has a predicted maximum bench press nearly 50 pounds higher than an athlete who can do only two 60-pound bench presses. Those predicted values are $\hat{y} = 63.5 + 1.49(2) = 66.5$ pounds at $x = 2$ and $\hat{y} = 63.5 + 1.49(35) = 115.7$ pounds at $x = 35$.

Insight

The slope of 1.49 is positive: As x increases, the predicted value $\hat{y}$ increases. The association is *positive*. When the association is *negative*, the predicted value $\hat{y}$ *decreases* as x increases. When the slope = 0, the regression line is horizontal.

Try Exercise 12.1

Example 4

Income and Education

Picture the Scenario

As described previously, suppose the regression line $\mu_y = -20{,}000 + 4000x$ models the relationship for the population of working adults in your hometown between x = number of years of education and the mean of y = annual income. This model tells us that income goes up as education does, on average, but how much variability is there? Suppose also that the conditional distribution of annual income at each value of x is modeled by a normal distribution, with standard deviation $\sigma = 13{,}000$.

Question to Explore

Use this regression model to describe the mean and variability around the mean for the conditional distributions of income at education values 12 and 16 years.

Think It Through

This regression model states that for workers with x years of education, their annual incomes have a normal distribution with a mean of $\mu_y = -20{,}000 + 4000x$ and a standard deviation of $\sigma = 13{,}000$. For those having a high school education ($x = 12$), the mean annual income is $\mu_y = -20{,}000 + 4000(12) = 28{,}000$ and the standard deviation is 13,000. Those with a college education ($x = 16$) have a mean annual income of $\mu_y = -20{,}000 + 4000(16) = 44{,}000$ and a standard deviation of 13,000.

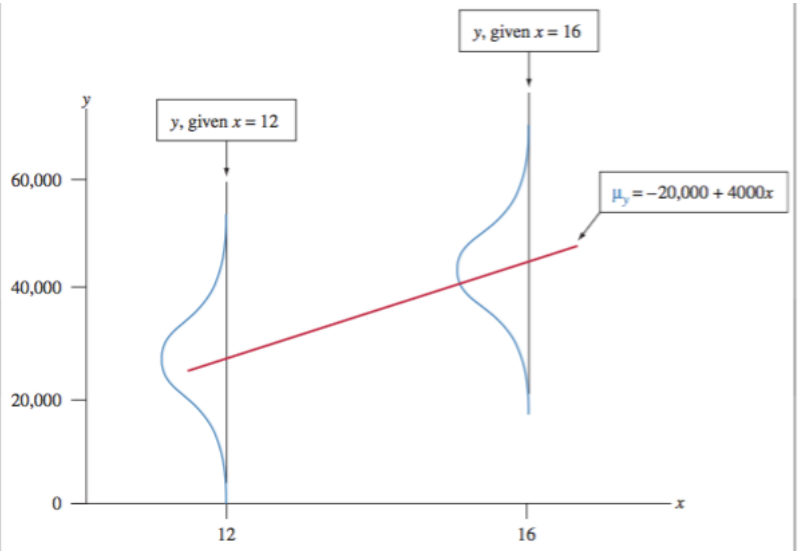
Since the conditional distributions are modeled as normal, the model predicts that nearly all of the y values fall within 3 standard deviations of the mean. For instance, for those with 16 years of education, $\mu_y = 44{,}000$ and $\sigma = 13{,}000$. Since

$$44{,}000 - 3(13{,}000) = 5000 \text{ and } 44{,}000 + 3(13{,}000) = 83{,}000$$

the model predicts that nearly all of their annual incomes fall between \$5000 and \$83,000.

Insight

Figure 12.4 portrays this regression model. It plots the regression equation and the conditional distributions of y = income at $x = 12$ years and at $x = 16$ years of education. We'll see how to use data to estimate the variability, as described by the standard deviation σ, later in the chapter.



▲ **Figure 12.4** The Regression Model $\mu_y = -20{,}000 + 4000x$, with $\sigma = 13{,}000$, Relating the Means of y = Annual Income to x = Years of Education. **Question** What do the bell-shaped curves around the line at $x = 12$ and at $x = 16$ represent?

Try Exercise 12.4

Example 5

Worldwide Internet Use

Picture the Scenario

The Twelve Countries data set on the text CD shows recent data for 12 countries on several variables. (*Source:* Data from www.internetworldstats.com and www.checkfacebook.com.) One of the variables in the file is Internet use, the percentage of adult residents who use the Internet. Which variables are strongly associated with Internet use? Let's consider three possibilities.

Unemployment rate: Total percentage of labor force unemployed

GDP: Gross domestic product, per capita, in thousands of U.S. dollars (a measure of a nation's economic development)

CO_2: Carbon dioxide emissions, per capita (a measure of air pollution)

Table 12.2 displays a correlation matrix for these four variables.

Table 12.2 Correlation Matrix for Internet Use, Unemployment Rate (%), Gross Domestic Product (GDP), and Carbon Dioxide (CO_2) Emissions

	Internet users (per 100 people)	Unemployment rate (%)	GDP per capita
Unemployment rate	0.238		
GDP per capita	0.938	0.163	
Carbon dioxide emissions	0.569	−0.107	0.740

Cell Contents: Pearson correlation

Source: Data from www.internetworldstats.com and www.checkfacebook.com.

Questions to Explore

a. Which variable has the strongest linear association with Internet use?

b. Which variable has the next strongest linear association with Internet use? Interpret.

Think It Through

a. The correlations with Internet use appear in the first column. The variable most strongly linearly associated with Internet use is the one having the largest correlation, in absolute value. This is GDP. Its correlation with Internet use is $r = 0.938$, a very strong positive association.

b. The variable Carbon Dioxide (CO_2) Emissions has the next strongest linear association with Internet use with correlation, $r = 0.569$. As CO_2 increases, there appears to be a moderate tendency for Internet use to increase.

Insight

We observe that the correlation ($r = 0.74$) between GDP and Carbon Dioxide Emissions indicates a strong positive association. This should not be surprising. Nations that are more economically advanced tend to have both higher GDP and CO_2. We have already observed that these two variables (GDP and CO_2) have a moderate to strong positive association with Internet use.

Try Exercise 12.12

Example 6

Predicting Strength

Picture the Scenario

For the female athlete strength study (Examples 1–3), x = number of 60-pound bench presses and y = maximum bench press had:

x: mean = 11.0, standard deviation = 7.1

y: mean = 79.9, standard deviation = 13.3 (both in pounds)

regression equation: $\hat{y} = 63.5 + 1.49x$.

Questions to Explore

a. Find the correlation r between these two variables.

b. Show that r does not change value if you measure y in kilograms.

Think It Through

a. The slope of the regression equation is $b = 1.49$. Since $s_x = 7.1$ and $s_y = 13.3$,

$$r = b\left(\frac{s_x}{s_y}\right) = 1.49\left(\frac{7.1}{13.3}\right) = 0.80.$$

The variables have a strong, positive association.

b. If y had been measured in kilograms, the y values would have been divided by 2.2, since 1 kg = 2.2 pounds. For instance, Subject 1 had $y = 80$ pounds, which is 80/2.2 = 36.4 kg. Likewise, the standard deviation s_y of 13.3 in pounds would have been divided by 2.2 to get 13.3/2.2 = 6.05 in kg. The slope of 1.49 would have been divided by 2.2, giving 0.68, since 1.49 pounds = 0.68 kg. Then

$$r = b(s_x/s_y) = 0.68(7.1/6.05) = 0.80.$$

The correlation is the same (0.80) if we measure y in pounds or in kilograms.

Insight

Now if we change units from kilograms to grams, s_y changes from 6.05 to 6050, b changes from 0.68 to 680, but again $r = 0.80$ because r does not depend on the units.

Try Exercise 12.15

Example 7

Tall Parents and Tall Children

Picture the Scenario

British scientist Francis Galton discovered the basic ideas of regression and correlation in the 1880s. He observed that very tall parents tended to have tall children, but on average not quite so tall. For instance, for all fathers with height 7 feet, their sons averaged 6 feet 5 inches when fully grown—taller than average, but not extremely tall. Likewise, for fathers with height 5 feet, perhaps their sons averaged 5 feet 5 inches—shorter than average, but not extremely short.

In his research, Galton accounted for gender height differences by multiplying each female height by 1.08, so heights of mothers and daughters had about the same mean as heights of fathers and sons. Then, for each son or daughter, he summarized their father's and mother's heights by parents' height = (father's height + mother's height)/2, the mean of their heights.

Question to Explore

Galton found a correlation of 0.5 between x = parents' height and y = child's height. How does his observation about very tall or very short parents with children who are not so very tall or so very short relate to the property about the correlation that a predicted value of y is relatively closer to its mean than x is to its mean?

Think It Through

From the property of r with $r = 0.5$, when x = parents' height is a certain number of standard deviations from its mean, then y = child's predicted height is *half* as many standard deviations from its mean. For example, if the parents' height is 2 standard deviations above the mean, the child is predicted to be 1 standard deviation above the mean (half as far, in relative terms, when $r = 0.5$). At $x = \bar{x} + 3s_x$, we predict $\hat{y} = \bar{y} + 1.5s_y$. In each case, on average a child's height is above the mean, but only half as far above the mean as their parent's height is above the parent's mean.

Insight

The correlation r is no greater than 1, in absolute value. So, a y value is predicted to be fewer standard deviations from its mean than x is from its mean.

Try Exercise 12.26

Example 8

The Placebo Effect

Picture the Scenario

A clinical trial admits subjects suffering from high blood cholesterol (over 225 mg/dl). The subjects are randomly assigned to take either a placebo or a drug being tested for reducing cholesterol levels. After the three-month study, the mean cholesterol level for subjects taking the drug drops from 270 to 230. However, the researchers are surprised to see that the mean cholesterol level for the placebo group also drops, from 270 to 250.

Question to Explore

Explain how this placebo effect could merely reflect regression toward the mean.

Think It Through

For a group of people, a person's cholesterol reading at one time would likely be positively correlated with their reading three months later. So, for all people who are not taking the drug, a subject with relatively high cholesterol at one time would also tend to have relatively high cholesterol three months later. By regression toward the mean, however, subjects who are relatively high at one time will, on average, be lower at a later time. So, if a study gives placebo to people with relatively high cholesterol (that is, in the right-hand tail of the blood cholesterol distribution), on average we expect their values three months later to be lower.

Insight

Regression toward the mean is pervasive. In sports, excellent performance tends to be followed by good, but less outstanding, performance. A football team that wins all its games in the first half of its schedule will probably not win all its games in the second half. A baseball player who hits 0.400 in the first month will probably not be hitting that high at the end of the season.

By contrast, the good news about regression toward the mean is that very poor performance tends to be followed by improved performance. If you got the worst score in your statistics class on the first exam, you probably did not do so poorly on the second exam (but you were probably still below the mean).

Try Exercise 12.23

Example 9

The Strength Study

Picture the Scenario

For the female athlete strength study, Example 6 showed that x = number of 60-pound bench presses and y = maximum bench press had a correlation of 0.80.

Question to Explore

Find and interpret r^2.

Think It Through

Since the correlation $r = 0.80$, $r^2 = (0.80)^2 = 0.64$. For predicting maximum bench press, the regression equation has 64% less error than $\bar{y}$ has. "Error" here refers to the summary given by the sum of squared prediction errors.

Insight

Since $r^2 = 0.64$ is quite far from 0, we can predict y much better using the regression equation than using $\bar{y}$. In this sense, the association is quite strong.

Try Exercise 12.16, part c

Example 10

High School GPA Predicting College GPA

Picture the Scenario

Consider the correlation between high school GPA and later performance in college measured by college GPA.

Question to Explore

For which group would these variables have a stronger correlation: All students who graduate from Harvard University this year, or all students who graduate from college somewhere in the United States this year?

Think It Through

The magnitude of the correlation depends on the variability in high school GPA. For Harvard students, the high school GPAs will concentrate narrowly at the upper end of the scale. So, the correlation would probably be weak. By contrast, for all students who finish college, high school GPA values would range from very low to very high. We would likely see a stronger correlation for them.

Insight

What reason explains this property of the correlation? Recall the formula, $r = b(s_x/s_y)$. When we use a much wider range of x values, s_x increases a lot. So, if the slope does not change much and if the variability of the y values is not much larger with the expanded sample, r will tend to increase because it is proportional in value to s_x.

Try Exercise 12.29

Example 11

60-Pound Strength and Bench Presses

Picture the Scenario

One purpose of the strength study introduced in Example 1 was to analyze whether x = number of times an athlete can lift a 60-pound bench press helps us predict y = maximum number of pounds the athlete can bench press. Table 12.4 shows the regression analysis for the 57 female athletes, with x denoted by BP_60 and y denoted by BP. The margin shows screen shots from the TI-83+/84.

Table 12.4 MINITAB Printout for Regression Analysis of y = Maximum Bench Press (BP) and x = Number of 60-Pound Bench Presses (BP_60)

```
The regression equation is BP = 63.5 + 1.49 BP_60

Predictor       Coef     SE Coef       T         P
Constant      63.537       1.956   32.48     0.000
BP_60         1.4911       0.150    9.96     0.000

R-Sq = 64.3%
```

Questions to Explore

a. Conduct a two-sided significance test of the null hypothesis of independence.

b. Report the P-value for the alternative hypothesis of a positive association, which is a sensible one for these variables. Interpret the results in context.

Think It Through

a. From Table 12.4, the prediction equation is $\hat{y} = 63.5 + 1.49x$. Here are the steps of a significance test of the null hypothesis of independence.

1. **Assumptions:** The scatterplot in Figure 12.1, shown again in the margin, revealed a linear trend, with scatter of points having similar variability (or spread) at different x values. The straight-line regression model $\mu_y = \alpha + \beta x$ seems appropriate. The 57 female athletes were *all* the female athletes at a particular high school. This was a convenience sample rather than a random sample of the population of female high school athletes. Although the goal was to make inferences about that population, inferences are tentative because the sample was not random.
2. **Hypotheses:** The null hypothesis that the variables are independent is H_0: $\beta = 0$. The two-sided alternative hypothesis of dependence is H_a: $\beta \neq 0$.
3. **Test statistic:** For the sample slope $b = 1.49$, Table 12.4 reports standard error, $se = 0.150$. This is listed under "SE Coef," in the row of the table for the predictor (BP_60). For testing H_0: $\beta = 0$, the t test statistic is

$$t = (b - 0)/se = 1.49/0.150 = 9.96.$$

This is extremely large. The sample slope falls nearly 10 standard errors above the null hypothesis value. The t test statistic appears

in Table 12.4 under the column labeled "T." The sample size equals $n = 57$, so $df = n - 2 = 55$.

4. **P-value:** The P-value, listed in Table 12.4 under the heading "P," is 0.000 rounded to three decimal places. Software reports the P-value for the two-sided alternative, H_a: $\beta \neq 0$. It is the two-tailed probability of "more extreme values" above 9.96 and below −9.96. See the margin figure.

5. **Conclusion:** If H_0 that the population slope β equals 0 were true, it would be extremely unusual to get a sample slope as far from 0 as $b = 1.49$. The P-value gives very strong evidence against H_0. We conclude that an association exists between the number of 60-pound bench presses and maximum bench press.

b. The one-sided alternative H_a: $\beta > 0$ predicts a positive association. For it, the P-value is halved, because it is then the right-tail probability of $t > 9.96$. This also equals 0.000, to three decimal places. On average, we infer that maximum bench press increases as the number of 60-pound bench presses increases.

Insight

In practice, studies must often rely on convenience samples. Results may be biased if the study subjects differ in an important way from those in the population of interest. Here, inference is reliable only to the extent that the sample is representative of the population of female high school athletes. This is a common problem with studies of this type, in which it would be difficult to arrange for a random sample of all subjects but a sample is conveniently available locally. We can place more faith in the inference if similar results occur in other studies.

The printout in Table 12.4 also contains a standard error and t statistic for testing that the population y-intercept α equals 0. This information is usually not of interest. Rarely is there any reason to test the hypothesis that a y-intercept equals 0.

Try Exercise 12.33, part a

Example 12

Estimating the Slope for Predicting Maximum Bench Press

Picture the Scenario

For the female athlete strength study, the sample regression equation is $\hat{y} = 63.5 + 1.49x$. From Table 12.4, the sample slope $b = 1.49$ has standard error $se = 0.150$.

Questions to Explore

a. Construct a 95% confidence interval for the population slope β.

b. What are the plausible values for the increase in maximum bench press, on average, for each additional 60-pound bench press that a female athlete can do?

Think It Through

a. For a 95% confidence interval, the $t_{.025}$ value for $df = n - 2 = 55$ is $t_{.025} = 2.00$. (If your software does not supply t-scores, you can find or approximate $t_{.025}$ from Table B.) The confidence interval is

$$b \pm t_{.025}(se) = 1.49 \pm 2.00(0.150),$$
$$\text{which is } 1.49 \pm 0.30$$
$$\text{or } (1.2, 1.8).$$

b. We can be 95% confident that the population slope β falls between 1.2 and 1.8. On average, the maximum bench press increases by between 1.2 and 1.8 pounds for each additional 60-pound bench press that an athlete can do.

Insight

Confidence intervals and two-sided significance tests about slopes are consistent: When a two-sided test has P-value below 0.05, casting doubt on $\beta = 0$, the 95% confidence interval for β does not contain 0.

Try Exercise 12.33, part b

Example 13

Detecting an Underachieving College Student

Picture the Scenario

Two of the variables in the Georgia Student Survey data file on the text CD are college GPA and high school GPA (variables CGPA and HSGPA). These were measured for a sample of 59 students at the University of Georgia. Identifying y = CGPA and x = HSGPA, we find $\hat{y} = 1.19 + 0.64x$.

Question to Explore

MINITAB highlights observations that have standardized residuals with absolute value larger than 2 in a table of "unusual observations." Table 12.6 shows this data. Interpret the results for observation 59.

Table 12.6 Observations with Large Standardized Residuals in Student GPA Regression Analysis, as Reported by MINITAB

Obs	HSGPA	CGPA	Fit	Residual	St Resid ← standardized residuals
14	3.30	2.60	3.29	−0.69	−2.26R
28	3.80	2.98	3.61	−0.63	−2.01R
59	3.60	2.50	3.48	−0.98	− 3.14R

R denotes an observation with a large standardized residual.

Think It Through

Observation 59 is a student who had high school GPA $x = 3.60$, college GPA $y = 2.50$, predicted college GPA $\hat{y} = 3.48$ (the "fit"), and residual $= y - \hat{y} = -0.98$. The reported standardized residual of -3.14 indicates that the residual is 3.14 standard errors below 0. This student's actual college GPA is quite far below what the regression line predicts.

Insight

Based on his or her high school GPA and predicted college GPA, this student with an actual college GPA of 2.50 seems to be an underachiever in college.

Try Exercise 12.43

Example 14

College GPA

Picture the Scenario

For the regression model of Example 13 predicting college GPA from high school GPA, Figure 12.9 is a MINITAB histogram of the standardized residuals. The margin shows a MINITAB boxplot of the standardized residuals.

Question to Explore

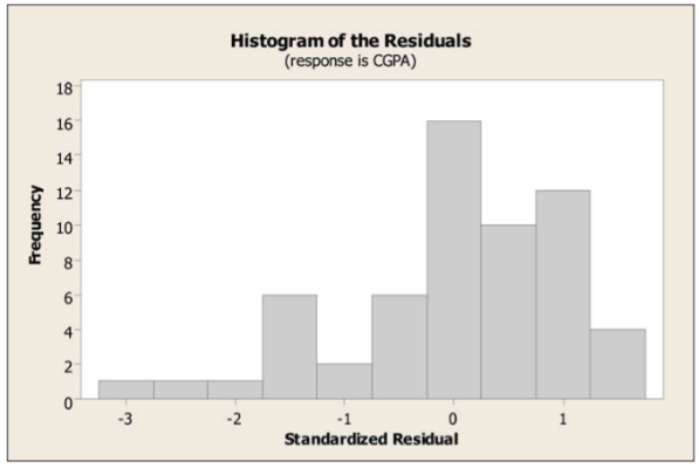
What does Figure 12.9 tell you?

Think It Through

The standardized residuals show some skew to the left. It may be that for a fixed value of high school GPA, college GPA tends to be skewed to the left, with some students doing much more poorly than the regression model would predict. The large negative standardized residual of -3.14 in Example 13, summarized by the left-most bar in Figure 12.9, may merely reflect skew in the distribution of college GPA. However, each of the three bars farthest to the left represents only a single observation, so this conclusion is tentative and requires a larger sample to check more thoroughly.

Insight

The sample size was 59. When n is not especially large, a graph like Figure 12.9 is an imprecise estimate of a corresponding graph for the population.



▲ **Figure 12.9** Histogram of Standardized Residuals for Regression Model Predicting College GPA. **Question** How many observations do the three left-most bars represent?

Although this graph shows some evidence of skew, much of this evidence is based on only three observations. This is not strong evidence that the conditional distribution is highly non-normal. In viewing such graphs, we need to be careful not to let a few observations influence our judgment too much. We're mainly looking for dramatic departures from the assumptions.

Try Exercise 12.45

Example 15

Variability of the Athletes' Strengths

Picture the Scenario

Let's return to the analysis of y = maximum bench press and x = number of 60-pound bench presses, for 57 female high school athletes. The prediction equation is $\hat{y} = 63.5 + 1.49x$. We'll see later in Table 12.8 that the residual sum of squares equals 3522.8.

Questions to Explore

a. Find the residual standard deviation of y.

b. Interpret the value you obtain in context at the sample mean value for x of 11.

Think It Through

a. Since $n = 57$, $df = n - 2 = 55$. The residual standard deviation of the y values is

$$s = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}} = \sqrt{\frac{3522.8}{55}} = \sqrt{64.1} = 8.0.$$

b. At any fixed value x of number of 60-pound bench presses, the model estimates that the maximum bench press values vary around a mean of $\hat{y} = 63.5 + 1.49x$ with a standard deviation of 8.0. See the figure in the margin.

At $x = 11$ (also the sample mean of the x values), the predicted maximum bench press is $\hat{y} = 63.5 + 1.49(11) = 79.9$. For female high school athletes who can do eleven 60-pound bench presses, we estimate that the maximum bench press values have a mean of about 80 pounds and a standard deviation of 8.0 pounds.

Insight

Why does the residual standard deviation ($s = 8.0$) differ from the standard deviation ($s_y = 13.3$) of the 57 y values in the sample? The reason is that s_y refers to the variability of *all* the y values around their mean, not just those at a fixed x value. That is, $s_y = 13.3$ describes variability about the overall mean of $\bar{y} = 80$ for *all* 57 high school female athletes, whereas $s = 8.0$ describes variability at a fixed x value such as $x = 11$. See the margin figure.

When the correlation is strong, at a fixed value of x we see less variability than the overall sample has. For instance, at the fixed value $x = 11$, we describe the variability in maximum bench press values by $s = 8.0$, whereas overall we describe variability in maximum bench press values by $s_y = 13.3$.

Try Exercise 12.49, part a

Example 16

Maximum Bench Press and Its Mean

Picture the Scenario

We've seen that the equation $\hat{y} = 63.5 + 1.49x$ predicts y = maximum bench press using x = number of 60-pound bench presses. For $x = 11$, Table 12.7 shows how MINITAB reports a confidence interval (CI) for the population mean of y and a prediction interval (PI) for a single y value. The predicted value, $\hat{y}$, is reported under the heading "Fit."

Table 12.7 MINITAB Output for Confidence Interval (CI) and Prediction Interval (PI) on Maximum Bench Press for Athletes Who Do Eleven 60-Pound Bench Presses before Fatigue

```
Predicted Values for New Observations

New Obs      Fit     SE Fit       95% CI            95% PI

    1      79.94     1.06    (77.81, 82.06)    (63.76, 96.12)

Values of Predictors for New Observations

New Obs     BP_60

    1       11.0
```

Questions to Explore

a. Using $\hat{y}$ and its *se* reported in Table 12.7, find and interpret a 95% *confidence interval* for the population mean of the maximum bench press values for all female high school athletes who can do $x = 11$ sixty-pound bench presses.

b. Report and interpret a 95% *prediction interval* for a single new observation on maximum bench press, for a randomly chosen female high school athlete with $x = 11$.

Think It Through

a. At $x = 11$, the predicted maximum bench press is $\hat{y} = 63.5 + 1.49(11) = 79.94$ pounds, the "fit." Table 12.7 reports a standard error for this estimate, $se = 1.06$, under the heading "SE Fit." With $n = 57$ we've seen that $df = n - 2 = 55$ and the t-score is $t_{.025} = 2.00$. So, the 95% confidence interval for the population mean of maximum bench press values at $x = 11$ is

$$\hat{y} \pm t_{.025}(se \text{ of } \hat{y}),$$
$$\text{which is } 79.94 \pm 2.00(1.06),$$
$$\text{or } 79.94 \pm 2.12,$$
$$\text{that is } (77.8, 82.1).$$

This is labeled as "95% CI" on the MINITAB printout. For all female high school athletes who can do eleven 60-pound bench presses, we are 95% confident that the mean of their maximum bench press values falls between about 78 and 82 pounds.

b. MINITAB reports the 95% prediction interval (63.8, 96.1) under the heading "95% PI." This predicts where maximum bench press y will fall for a randomly chosen female high school athlete having $x = 11$. Equivalently, this refers to where 95% of the corresponding population values fall. For all female high school athletes who can do eleven 60-pound bench presses, we predict that 95% of them have maximum bench press between about 64 and 96 pounds. Look at Figure 12.1, reproduced in the margin. Locate $x = 11$. Of all possible data points at that x value, we predict that 95% of them would fall between about 64 and 96.

Insight

The 95% prediction interval (63.8, 96.1) predicts the maximum bench press y for a randomly chosen female high school athlete having $x = 11$. The 95% confidence interval (77.8, 82.1) estimates the mean of such y values for all female high school athletes having $x = 11$. The prediction interval for a single observation y is much wider than the confidence interval for the mean of y. In other words, you can estimate a population mean more precisely than you can predict a single observation.

Try Exercise 12.47

Example 17

Growth in Population Size

Picture the Scenario

The population size of the United States has been growing rapidly in recent years, much of it due to immigration. According to the 2010 census, the population size was about 309 million on April 1, 2010.

Questions to Explore

a. Suppose that the rate of growth after 2010 is 2% a year. That is, the population is 2% larger at the end of each year than it was at the beginning of the year. Find the population size after (i) 1 year, (ii) 2 years, and (iii) 10 years.

b. Give a formula for the population size in terms of x = number of years since 2010.

Think It Through

a. With a 2% growth rate, the population size one year after 2010 is $309 \times 1.02 = 315.2$ million. That is, increasing the population size by 2% corresponds to multiplying it by 1.02. The population size after 2 years is 2% higher than this, or

$$315.2 \times 1.02 = (309 \times 1.02) \times 1.02 = 309 \times (1.02)^2 = 321.5 \text{ million.}$$

After 10 years, the population size is $309 \times 1.02 \times 1.02 \times 1.02 \times \ldots \times 1.02 = 309 \times (1.02)^{10} = 376.7$ million.

b. Can you see the pattern? For each additional year, we mulitply by another factor of 1.02. After x years, the population size is 309×1.02^x million.

Insight

The population size formula, 309×1.02^x, is called **exponential growth**. Plotted, the response goes up faster than a straight line. See the margin figure. The amount of change in y per unit change in x increases as x increases. After 100 years (that is, in the year 2110), taking $x = 100$, population size $= 309 \times (1.02)^{100} = 2238.6$ million, more than 2.2 billion people!

Try Exercise 12.56

Example 18

Explosion in Number of Facebook Users

Picture the Scenario

Table 12.9 shows the number of people (in millions) worldwide using Facebook between 2004 and 2011. Figure 12.12 plots these values. They increase over time, and the amount of increase from one year to the next seems to itself increase over time.

Table 12.9 also shows the logarithm of the Facebook-user counts, using base-10 logs. Figure 12.13 plots these log values over time. They appear to grow approximately linearly. In fact, the correlation between the log of the population size and the date is 0.985, a very strong linear association. This suggests that growth in Facebook users over this time period was approximately exponential.

Question to Explore

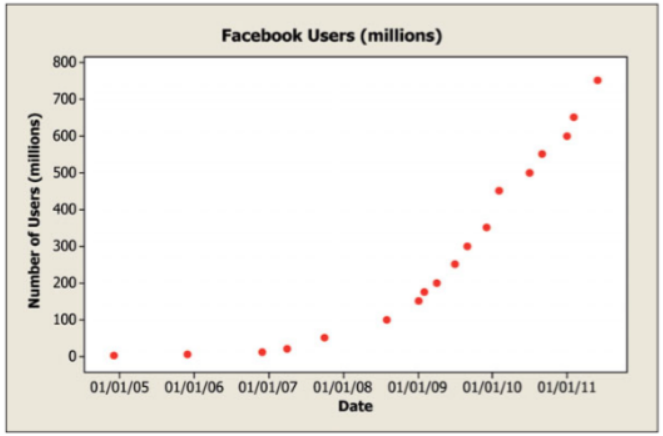
Let x denote the number of days since December 1, 2004. That is, December 1, 2004 is $x = 0$, December 1, 2005 is $x = 365$, and so forth up to June 1, 2011, which is $x = 2373$. Software provides the exponential regression model fitted to y = number of Facebook users and x gives

$$\hat{y} = 1.9559 \times 1.00275^{x}.$$

Table 12.9 Number of Facebook Users Worldwide (in Millions)

Date	Number of Days Since December 1, 2004 x	Number of Users (in millions) y	Log Number Users Log(y)	Predicted Number $\hat{y}$
12/01/04	0	1	0	1.9559
12/01/05	365	5.5	0.740363	5.329316
12/01/06	730	12	1.079181	14.52099
04/01/07	851	20	1.30103	20.24461
10/01/07	1034	50	1.69897	33.46325
08/01/08	1339	100	2	77.32727
01/01/09	1492	150	2.176091	117.7094
02/01/09	1523	175	2.243038	128.1693
04/01/09	1582	200	2.30103	150.7133
07/01/09	1673	250	2.39794	193.5016
09/01/09	1735	300	2.477121	229.4194
12/01/09	1826	350	2.544068	294.5529
02/01/10	1888	450	2.653213	349.2278
07/01/10	2038	500	2.69897	527.2413
09/01/10	2100	550	2.740363	625.1079
01/01/11	2222	600	2.778151	873.8982
02/01/11	2253	650	2.812913	951.5544
06/01/11	2373	750	2.875061	1322.983

Source: Data from Facebook User Growth Chart—2004–2011 by Ben Foster (www.benphoster.com/facebook_user_growth_chart_2004_2010/).



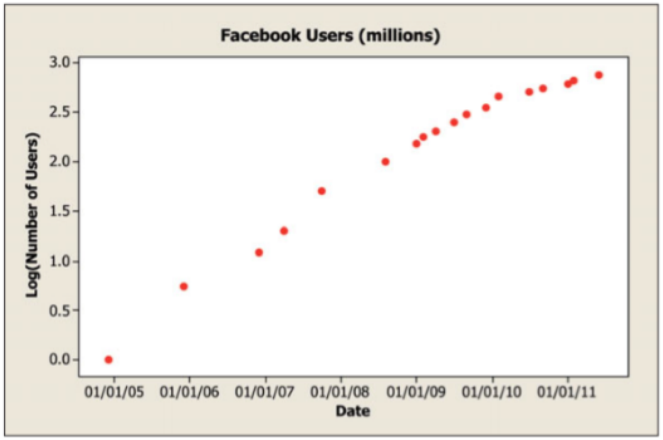
▲ **Figure 12.12** Plot of Number of Facebook Users (millions) from December 2004 to June 2011. *Source:* Graph by Ben Foster (twitter.com/benphoster) and updated at benphoster.com/facebookgrowth.

What does this equation predict for the number of Facebook users on Dec. 1, 2004, Dec. 1, 2007 (1095 days after Dec. 1, 2004), and on Dec. 1, 2015 (4017 days after Dec. 1, 2004)?

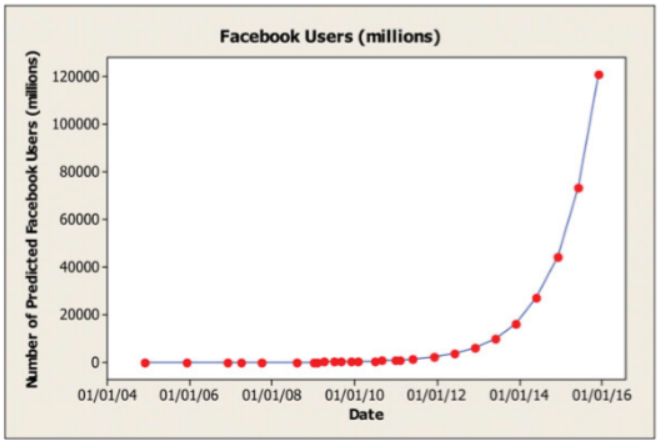
Think It Through

On December 1, 2004, $x = 0$, so the predicted number of Facebook users is $\hat{y} = 1.9559 \times 1.00275^{0} = 1.9559$ million users.

For the day December 1, 2007, $x = 1095$ days, and the predicted number is $\hat{y} = 1.9559 \times 1.00275^{1095} = 39.566$ million users.



▲ **Figure 12.13** Plot of Log of Number of Facebook Users Between 2004 and 2011. When the log of the response has an approximate straight-line relationship with the explanatory variable, the exponential regression model is appropriate.



▲ **Figure 12.14** Plot of Predicted Number of Facebook Users Between 2004 and 2011. The values after June 2011 are extrapolations beyond scope of data.

December 1, 2015 is 4017 days after December 1, 2004, and the prediction is $\hat{y} = 1.9559 \times 1.00275^{4017} = 120{,}866.1428$ million users, that is, almost 120 billion people. You should be skeptical of this prediction because the world population size at the end of 2010 is about 6.8 billion people and projected to be between 7 and 8 billion people by 2015. It is impossible to have more Facebook users than people on the planet.

Insight

Table 12.9 shows the predicted value for each date recorded between 2004 and 2011. Comparing the observed values in Table 12.9 to the predicted values for this model, we see that the model fits the data pretty well. However, there is some indication that the actual growth was slowing a bit in 2010, as the observation for September 1, 2010 (550 million users), was quite a bit less than the predicted value (625.1079 million users). Figure 12.14 extends the graph to show also data for the years between 2011 and 2020. The prediction model extrapolates the data after June 2011 using the exponential model. It's extremely unlikely that the world population will be as high as the prediction for Facebook users. This is another example of the danger of extrapolating a regression model beyond predictor values for which we have data.

Try Exercise 12.58