axis) increases by one unit. A *negative slope* indicates a decrease in one variable as the other increases, just as a negative correlation does.

For example, Figure 10.5 shows the (deterministic) relationship between y = temperature in Fahrenheit and x = temperature in Celsius. The equation for the relationship is

$$y = 32 + 1.8x$$

The intercept, 32, is the temperature in Fahrenheit when the Celsius temperature is zero. The slope, 1.8, is the amount by which Fahrenheit temperature increases when Celsius temperature increases by one unit.

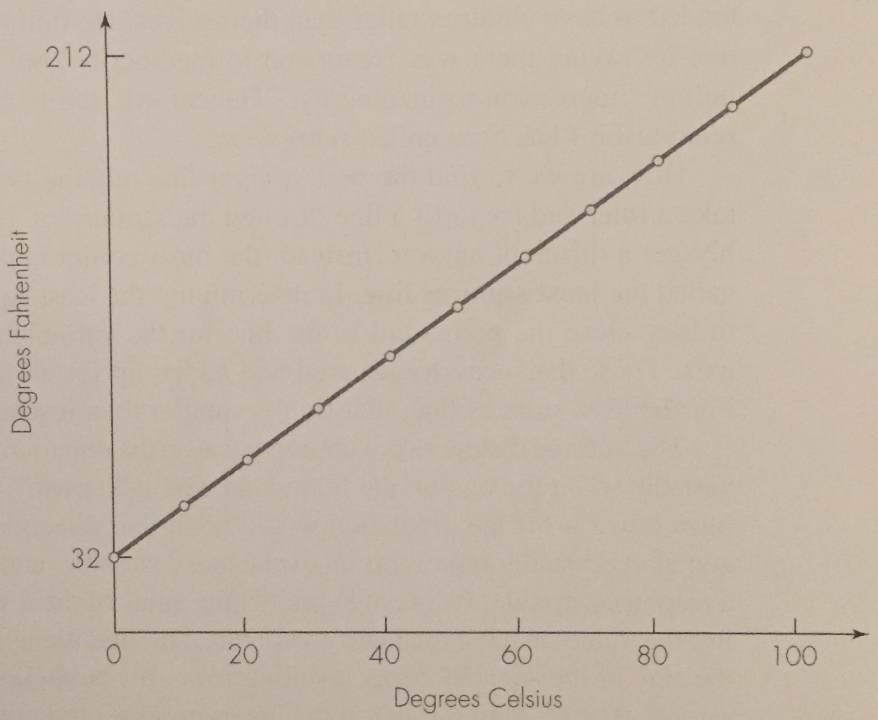

Figure 10.5
A straight line with intercept of 32 and slope of 1.8

Finding and Using the Regression Equation

There are many calculators, computer programs and websites that will find the intercept and slope of the least squares line for you. Once you have the intercept and slope, you can combine them into the regression equation as follows:

$$y = \textit{intercept} + \textit{slope} \times \textit{(x-value)}$$

However, remember that in a statistical relationship knowing the value of x does not give you a precise value for y. Instead, the equation can be used for the following two purposes:

- To predict the value of y in the future, when only the value of x (and not y) is known. In this case, the equation is written as:

$$\textit{Predicted } y = \textit{intercept} + \textit{slope} \times \textit{(x-value)}$$

- To estimate the average value of y for a particular value of x. In this case, the equation is written as:

$$\textit{Estimated } y = \textit{intercept} + \textit{slope} \times (\textit{x-value})$$

For instance, we might want to use x = verbal SAT to predict y = college GPA using the data exhibited in Figure 9.5 on page 186. The regression equation for doing so, found using the data in Figure 9.5, is:

$$\textit{Predicted GPA} = 0.539 + (0.00362)(\textit{verbal SAT})$$

Using this equation, someone with a verbal SAT score of 600 would be predicted to have a GPA of:

$$\textit{Predicted GPA} = 0.539 + (0.00362)(600) = 2.71$$

This is also what we would estimate as the average (mean) GPA for all individuals with a verbal SAT score of 600. But, remember that it is not going to be the exact GPA for *each* of those individuals, because the relationship is statistical and not deterministic. Some individuals will have higher GPAs, and some will have lower GPAs. But the average will be close to 2.71.

Using Excel to Find the Intercept and Slope of the Least Squares Regression Line

Here is how you find the intercept and slope in Excel:

1. List the y values in one column. Let's call the cells used "Array 1." For instance, if you have 10 individuals and use Column A, rows 1 to 10 for their y values, then Array 1 is A1:A10.
2. List the x values in the same order in another column. Let's call the cells used "Array 2." For instance, Array 2 might be B1:B10, representing the first 10 rows of column B. Array 1 and Array 2 must be the same length.
3. The Excel function INTERCEPT(Array 1, Array 2) gives the value of the intercept.
4. The Excel function SLOPE(Array 1, Array 2) gives the value of the slope.

CAUTION! Note that the first array listed contains the y-values, and the second one contains the x-values. Usually we think of the x values as being listed first, so you have to remember that the order required by Excel is to list the y values first.

EXAMPLE 10.7 Husbands' and Wives' Ages, Revisited

Figure 10.6 shows the same scatterplot as Figure 10.1, relating ages of husbands and wives, except that now we have added the regression line. This line minimizes the sum of the squared vertical distances between the line and the husbands' actual

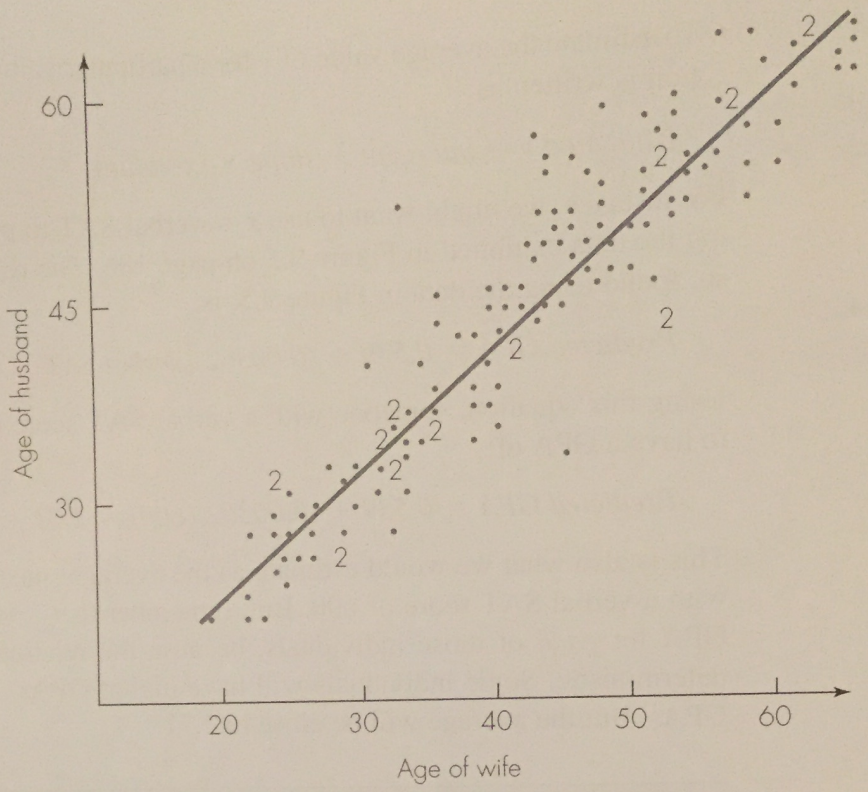

Figure 10.6
Scatterplot and regression line for British husbands' and wives' ages
Source: Hand et al., 1994.

ages. The regression equation for the line shown in Figure 10.6, relating husbands' and wives' ages, is

$$\text{Predicted } y = 3.6 + .97x$$

or, equivalently,

$$\text{Predicted husband's age} = 3.6 + (.97)(\text{wife's age})$$

Notice that the intercept of 3.6 does not have any meaning in this example. It would be the predicted age of the husband of a woman whose age is 0. But obviously that's not a possible wife's age. The slope does have a reasonable interpretation. For every year of difference in two wives' ages, there is a difference of about .97 years in their husbands' ages, close to 1 year. For instance, if two women are 10 years apart in age, their husbands can be expected to be about $(.97) \times 10 = 9.7$ years apart in age.

Let's use the equation to predict husband's age at various wife's ages.

Wife's Age	Predicted Age of Husband
20 years	3.6 + (.97)(20) = 23.0 years
25 years	3.6 + (.97)(25) = 27.9 years
40 years	3.6 + (.97)(40) = 42.4 years
55 years	3.6 + (.97)(55) = 57.0 years

This table shows that for the range of ages in the sample, husbands tend to be 2 to 3 years older than their wives, on average. The older the couple, the smaller the gap in their ages. Remember that with statistical relationships, we are determining what

happens to the average and not to any given individual. Thus, although most couples won't fit the pattern given by the regression line exactly, it does show us one way to represent the average relationship for the whole group. ■

Investigating Long-Term Trends in Time Series

In Chapter 9, we learned that a *time series* is a measurement variable recorded at evenly spaced intervals over time. Recall that one of the four components that help explain data in a time series is *long-term trend.* If the trend is linear, we can estimate the trend by finding a regression line, with time period as the explanatory variable and the measurement in the time series as the response variable. We can then remove the trend to enable us to see what other interesting features exist in the series. When we remove the linear trend in a time series, the result is, aptly enough, called a **detrended time series**. Let's revisit jeans sales in Britain, and see what we can learn about the trend.

EXAMPLE 10.8 **Jeans Sales in Britain, Revisited**

Figure 9.6 illustrated a time series of sales of jeans in Britain from 1980 to 1984. Figure 10.7 shows the same time series plot, with the regression line for month versus sales superimposed. Month 1 is January 1980, and month 60 is December 1984. The equation for the line is:

$$Sales = 1880 + 6.62\,(month)$$

If we were to try to forecast sales for January 1985, the first month that is not included in the series, we would use month = 61 and solve the equation for "Sales." The resulting

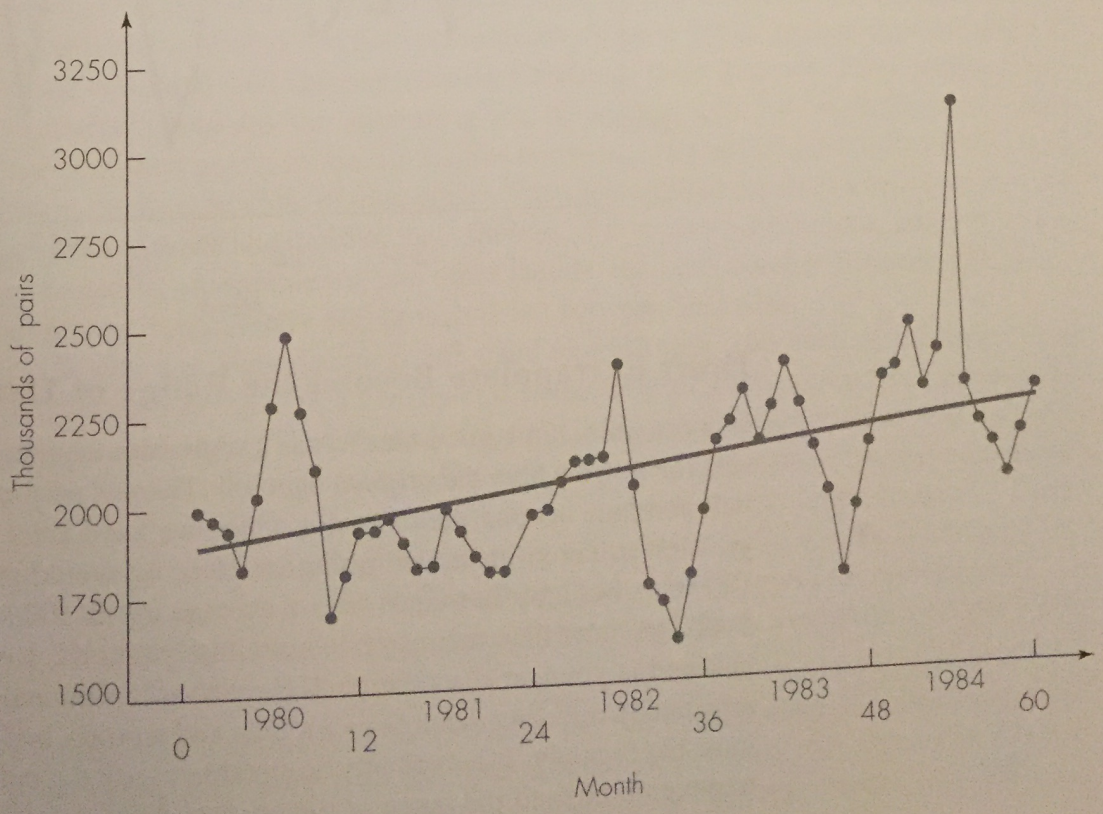

Figure 10.7 Jeans sales from Figure 9.6 with regression line showing linear trend.

value is 1880 + 6.62(61) = 2284 thousand pairs of jeans. Actual sales for January 1985 were 2137 thousand pairs. Our prediction is not far off, given that, overall, the data range from about 1600 to 3100 thousand pairs. One reason the actual value may be slightly lower than the predicted value is that sales tend to be lower during the winter months. Remember that seasonal components are another of the factors that affect values in a time series, and in this example, sales tend to be low in January for all years.

The regression line indicates that the trend, on average, is that sales increase by about 6.62 units per month. Because the units represent thousands of pairs, the actual increase is about 6620 pairs per month. Figure 10.8 presents the time series for jeans sales with the trend removed. Compare Figure 10.8 with Figure 10.7. Notice that the fluctuations remaining in Figure 10.8 are similar in character to those in Figure 10.7, but the upward trend is gone. ■

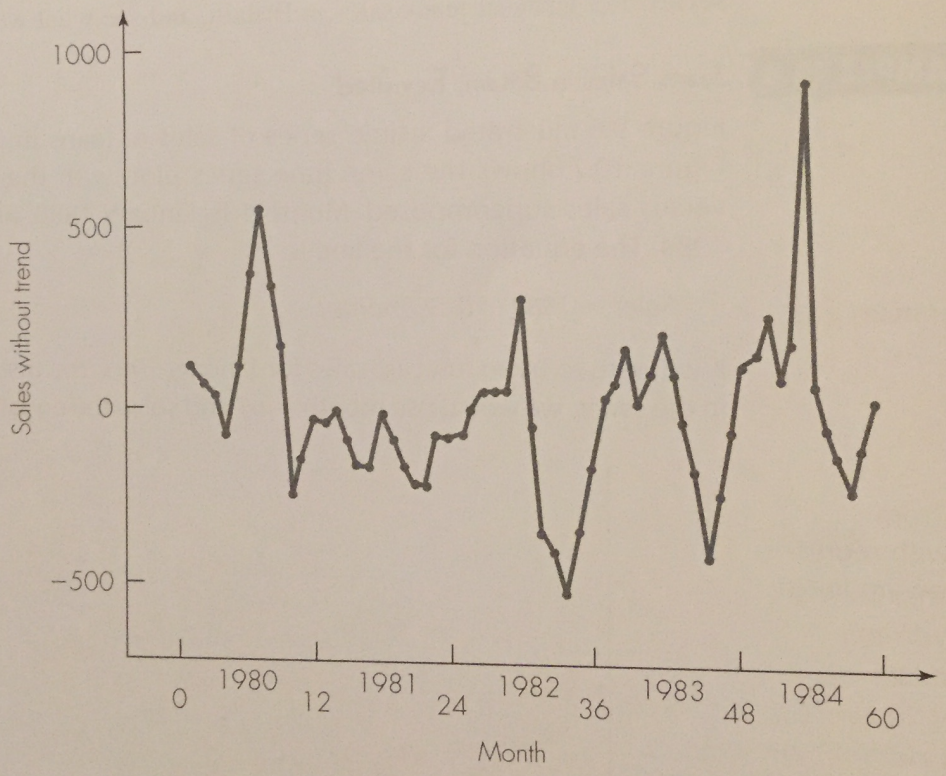

Figure 10.8
Detrended time series of jeans sales (linear trend removed).

Don't Extrapolate Beyond the Range of Data Values

It is generally not a good idea to use a regression equation to predict values far outside the range where the original data fell. There is no guarantee that the relationship will continue beyond the range for which we have data. For example, using the regression equation illustrated in Figure 10.6, we would predict that women who are 100 years old have husbands whose average age is 100.6 years. But women tend to live to be older than men, so it is more likely that if a woman is married at 100, her husband is younger than she is. The relationship for much older couples would be affected by differing death rates for men and women, and a different equation would most likely apply. It is typically acceptable to use the equation only for a minor extrapolation beyond the range of the original data.

A Final Cautionary Note

It is easy to be misled by inappropriate interpretations and uses of correlation and regression. In the next chapter, we examine how that can happen, and how you can avoid it.

CASE STUDY 10.1

Are Attitudes about Love and Romance Hereditary?

SOURCE: Waller and Shaver (September 1994).

Are you the jealous type? Do you think of love and relationships as a practical matter? Which of the following two statements better describes how you are likely to fall in love?

> My lover and I were attracted to each other immediately after we first met.
>
> It is hard for me to say exactly when our friendship turned into love.

If the first statement is more likely to describe you, you would probably score high on what psychologists call the *Eros* dimension of love, characteristic of those who "place considerable value on love and passion, are self-confident, enjoy intimacy and self-disclosure, and fall in love fairly quickly" (Waller and Shaver, 1994, p.268). However, if you identify more with the second statement, you would probably score higher on the *Storge* dimension, characteristic of those who "value close friendship, companionship, and reliable affection" (p. 268). Whatever your beliefs about love and romance, do you think they are partially inherited, or are they completely due to social and environmental influences?

Psychologists Niels Waller and Philip Shaver set out to answer the question of whether feelings about love and romance are partially genetic, as are most other personality traits. Waller and Shaver studied the love styles of 890 adult twins and 172 single twins and their spouses from the California Twin Registry. They compared the similarities between the answers given by monozygotic twins (MZ), who share 100% of their genes, to the similarities between those of dizygotic twins (DZ), who share, on average, 50% of their genes. They also studied the similarities between the answers of twins and those of their spouses. If love styles are genetic, rather than determined by environmental and other factors, then the matches between MZ twins should be substantially higher than those between DZ twins.

Waller and Shaver studied 345 pairs of MZ twins, 100 pairs of DZ twins, and 172 spouse pairs (that is, a twin and his or her spouse). Each person filled out a questionnaire called the "Love Attitudes Scale" (LAS), which asked them to read 42 statements like the two given earlier. For each statement, respondents assigned a ranking from 1 to 5, where 1 meant "strongly agree" and 5 meant "strongly disagree." There were seven questions related to each of six love styles, with a score determined for each person on each love style. Therefore, there were six scores for each person.

In addition to the two styles already described (Eros and Storge), scores were generated for the following four:

- *Ludus* characterizes those who "value the fun and excitement of romantic relationships, especially with multiple alternative partners; they generally are not interested in mutual self-disclosure, intimacy, or 'getting serious' " (p. 268).

(Continued)

- *Pragma* types are "pragmatic, entering a relationship only if it meets certain practical criteria" (p. 269).
- *Mania* types "are desperate and conflicted about love. They yearn intensely for love but then experience it as a source of pain, a cause of jealousy, a reason for insomnia" (p. 269).
- Those who score high on *Agape* "are oriented more toward what they can give to, rather than receive from, a romantic partner. Agape is a selfless, almost spiritual form of love" (p. 269).

For each type of love style, and for each of the three types of pairs (MZ twins, DZ twins, and spouses), the researchers computed a correlation. The results are shown in Table 10.1. (They first removed effects due to age and gender, so the correlations are not due to a relationship between love styles and age or gender.) Notice that the correlations for the MZ twins are lower than they are for the DZ twins for two love styles, and just slightly higher for the other four styles. This is in contrast to most other personality traits. For comparison purposes, three such traits are also shown in Table 10.1. Notice that for those traits, the correlations are much higher for the MZ twins, indicating a substantial hereditary component. Regarding the findings for love styles, Waller and Shaver conclude:

> *This surprising, and very unusual, finding suggests that genes are not important determinants of attitudes toward romantic love. Rather, the common environment appears to play the cardinal role in shaping familial resemblance on these dimensions.* (p. 271) ■

TABLE 10.1 Correlations for Love Styles and for Some Personality Traits

	Correlation		
	Monozygotic Twins	Dizygotic Twins	Spouses
Love Style			
Eros	.16	.14	.36
Ludus	.18	.30	.08
Storge	.18	.12	.22
Pragma	.40	.32	.29
Mania	.35	.27	−.01
Agape	.30	.37	.28
Personality Trait			
Well-being	.38	.13	.04
Achievement	.43	.16	.08
Social closeness	.38	.01	−.04

Source: Waller and Shaver, September 1994.

CASE STUDY 10.2

A Weighty Issue: Women Want Less, Men Want More

Do you like your weight? Let me guess . . . If you're male and under about 175 pounds, you probably want to weigh the same or more than you do. If you're female, no matter what you weigh, you probably want to weigh the same or less. Those were the results uncovered in a large statistics class (119 females and 63 males) when students were asked to give their actual and their ideal weights.

Figure 10.9 shows a scatterplot of ideal versus actual weight for the females, and Figure 10.10 (next page) is the same plot for the males. Each point represents one student, whose ideal weight can be read on the vertical axis and actual weight can be read on the horizontal axis. What is the relationship between ideal and actual weight, on average, for men and for women?

First, notice that if everyone were at their ideal weight, all points would fall on a line with the equation

$$\text{ideal} = \text{actual}$$

That line is drawn in each figure. Most of the women fall below that line, indicating that their ideal weight is below their actual weight. The situation is not as clear for the men, but a pattern is still evident. Most of those weighing under 175 pounds fall on or above the line (would prefer to weigh the same or more than they do), and most of those weighing over 175 pounds fall on or below the line (would prefer to weigh the same or less than they do).

The regression lines are also shown on each scatterplot. The regression equations are:

$$\text{Women: ideal} = 43.9 + 0.6\ \text{actual}$$
$$\text{Men: ideal} = 52.5 + 0.7\ \text{actual}$$

(Continued)

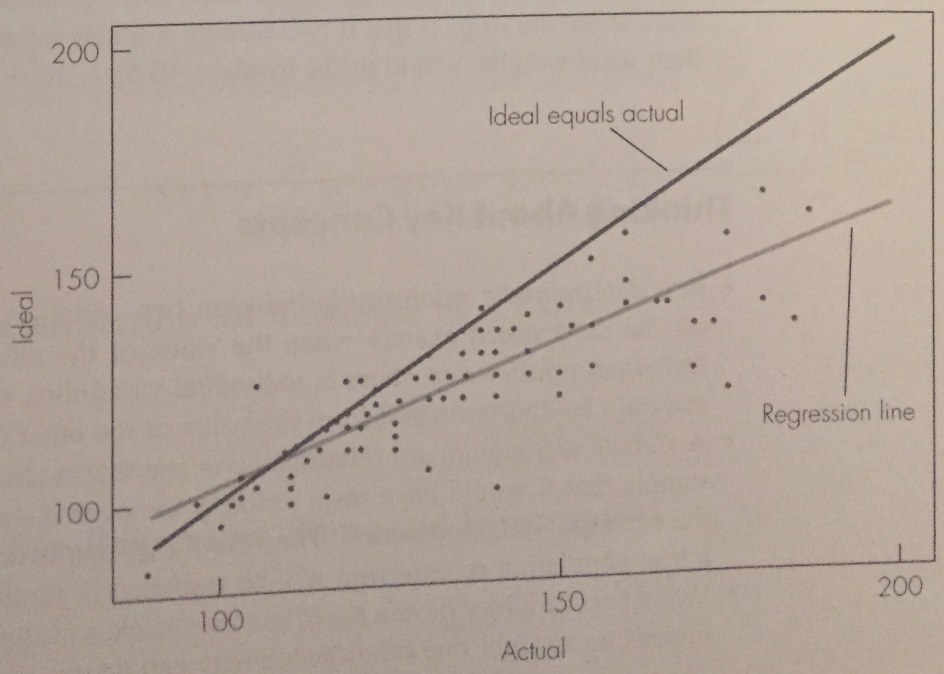

Figure 10.9
Ideal versus actual weight for females

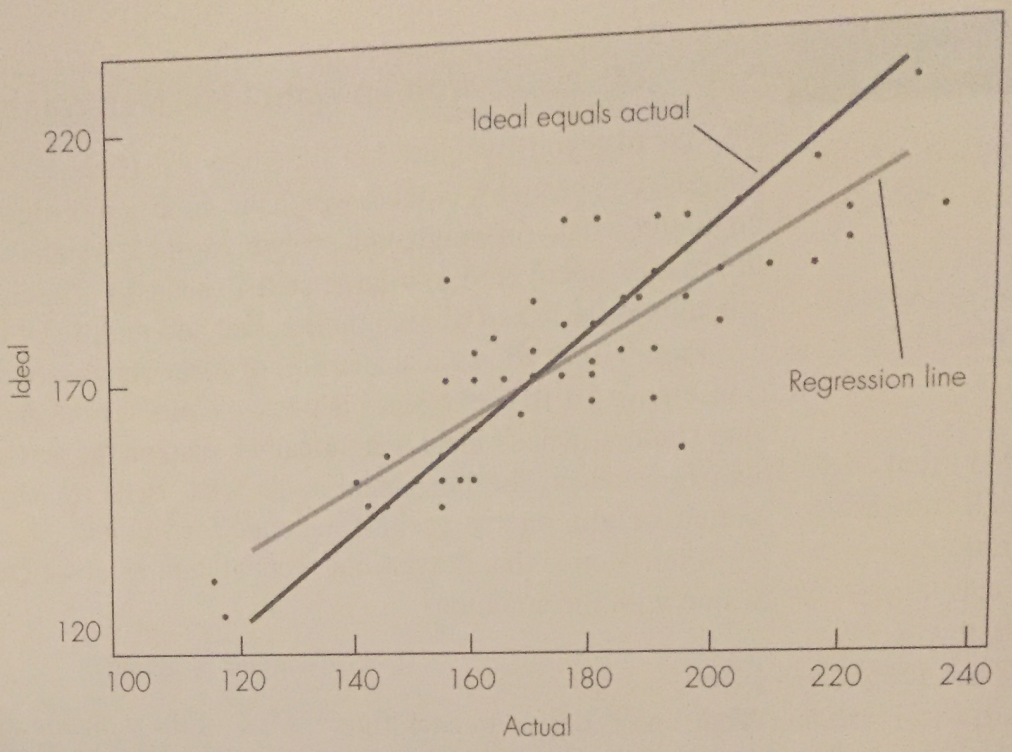

Figure 10.10
Ideal versus actual weight for males

These equations have several interesting features, which, remember, summarize the relationship between ideal and average weight for the aggregate, not for each individual:

- The weight for which *ideal* = *actual* is about 110 pounds for women and 175 pounds for men. Below those weights, actual weight is less than desired; above them, actual weight is more than desired.
- The slopes represent the increase in ideal weight for each 1-pound increase in actual weight. Thus, every 10 pounds of additional weight indicates an increase of only 6 pounds in ideal weight for women and 7 pounds for men. Another way to think about the slope is that if two women's actual weights differed by 10 pounds, their ideal weights would differ by about $(0.6) \times 10 = 6$ pounds. ■

Thinking About Key Concepts

- In a *deterministic relationship* between two variables, the value of one variable can be determined exactly when the value of the other one is known. But in a *statistical relationship*, there is individual variability, so the value of one variable can only be approximated from the value of the other one.
- A *statistically significant* relationship is one that is strong enough in the observed sample that it would have been unlikely to occur if there were no relationship in the corresponding population. The size of the sample influences the likelihood that a true population relationship will be statistically significant in the sample.
- When two variables have a *positive correlation*, a scatterplot would show a general increasing straight line relationship between them.

- When two variables have a *negative correlation,* a scatterplot would show a general decreasing straight line relationship between them.
- A *regression line* is a straight line placed through the points on a scatterplot showing the average relationship between the two variables, and a *regression equation* is the equation for that line. This equation can be used to predict values of y when the value of x is known. It can also be used to estimate the average value of y for a given value of x.
- A regression equation should never be used to *extrapolate* values far beyond the range of the data used to create the equation because the relationship may not stay the same.
- In a *detrended time series* plot, the linear trend has been removed, which allows the other components to be more readily seen.

Focus On Formulas

The Data

n pairs of observations, (x_i, y_i), $i = 1, 2, \ldots, n$, where x_i is plotted on the horizontal axis and y_i on the vertical axis.

Summaries of the Data, Useful for Correlation and Regression

$$\text{SSX} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$\text{SSY} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$\text{SXY} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

Correlation for a Sample of n Pairs

$$r = \frac{\text{SXY}}{\sqrt{\text{SSX}}\sqrt{\text{SSY}}}$$

The Regression Slope and Intercept

$$\text{slope} = b = \frac{\text{SXY}}{\text{SSX}}$$

$$\text{intercept} = a = \bar{y} - b\bar{x}$$