AN INTRODUCTION TO STATISTICS

WITH

DATA ANALYSIS

by

SHELLEY RASMUSSEN

Department of Mathematical Sciences
Olney 428T
University of Massachusetts/Lowell
Lowell, MA 01854

Originally published by Brooks/Cole Publishing Company,
Division of Wadsworth, Inc.

ISBN 0-534-13578-1

Shelley_Rasmussen@uml.edu

Minitab is a statistical package, a computer program that performs many statistical procedures. The versions of Minitab now available for use on personal computers are menu-driven and much easier to use than the main-frame version originally discussed in this text. Those sections are not included in this online edition of the text. At this time, the most recent version is Minitab 15, available at very reasonable prices for purchase or rental from:

www.e-academy.com/minitab

System Requirements

Processor:	PC with a 1 GHz 32- or 64-bit processor
Memory:	512 MB or more of available RAM
Disk Space:	125 MB free space available
Operating System:	Microsoft Windows 2000, XP, or Vista.
Display:	A display capable of 1024 X 768 or higher resolution
Software:	Adobe Acrobat Reader 5.0 or higher for Meet Minitab

CHAPTER 15

Correlation, Regression, and the Method of Least Squares

IN THIS CHAPTER

Correlation coefficient
Rank correlation coefficient
Method of least squares
Simple linear regression
Comparing the standard deviation line with the least squares line
Multiple regression

We will now look at ways to study relationships between quantitative variables. For instance: What is the relationship between height and weight in young children? How does income vary with education? Does blood pressure depend on age in adults? What is the relationship between advertising expenditures and sales?

One possible relationship between two quantitative variables is linear: A scatterplot of the two variables looks roughly like a straight line. The correlation coefficient is a measure of the extent of *linear* association between two quantitative variables, as we see in Section 15-1. A parametric test of independence of two quantitative variables is discussed in Section 15-2. In Section 15-3 we consider a correlation coefficient based on ranks, and discuss a nonparametric test of independence of two quantitative variables.

We may want to model one quantitative variable as a straight-line function of another. The method of least squares allows us to fit a straight line to a set of points in a scatterplot. Finding such a straight line and testing hypotheses about the model is called *simple linear regression,* discussed in Section 15-4. The relationship between linear correlation and the least squares line found in simple linear regression is the subject of Section 15-5. Also included are examples of how to interpret the phrase *regression toward the mean.*

Section 15-6 provides a very brief introduction to multiple regression. In multiple regression, we try to model a quantitative variable as a function of other variables.

Let's begin with the linear correlation coefficient. We use the correlation coefficient to measure linear association between two quantitative variables.

15-1 The Linear Correlation Coefficient

The linear correlation coefficient is a descriptive statistic. We use it to measure linear association between two quantitative variables, as a tool in data analysis. In this section we are concerned with the correlation coefficient only as a descriptive statistic; we make no inferences based on it. Therefore, it is appropriate to return to the World Bank data set to provide examples.

Values of four World Bank indicators are shown in Table 15-1 for four high-income, oil-exporting nations (World Bank, 1987). A scatterplot of birth rate versus per capita gross national product is shown in Figure 15-1 for these four countries; the points lie close to a straight line with negative slope. Figure 15-2 is a plot of life expectancy versus per capita gross national product. The association between the variables in this graph is less strongly linear; the points do not appear to lie as close to a straight line as the ones in Figure 15-1. Figure 15-3 shows a scatterplot of calorie supply versus per capita gross national product; there is no linear relationship apparent in this graph.

We use the linear correlation coefficient to measure the extent of *linear* association between two quantitative variables. Let's define the linear correla-

TABLE 15-1 Values of 1985 per capita gross national product, 1985 birth rate per 1,000 population, 1985 life expectancy at birth, and 1985 daily calorie supply per capita are listed for four high-income, oil-exporting nations.

Country	Gross national product	Birth rate	Life expectancy	Calorie supply
Libya	$7,170	45	60	3,612
Saudi Arabia	$8,850	42	62	3,128
Kuwait	$14,480	34	72	3,138
United Arab Emirates	$19,270	30	70	3,625

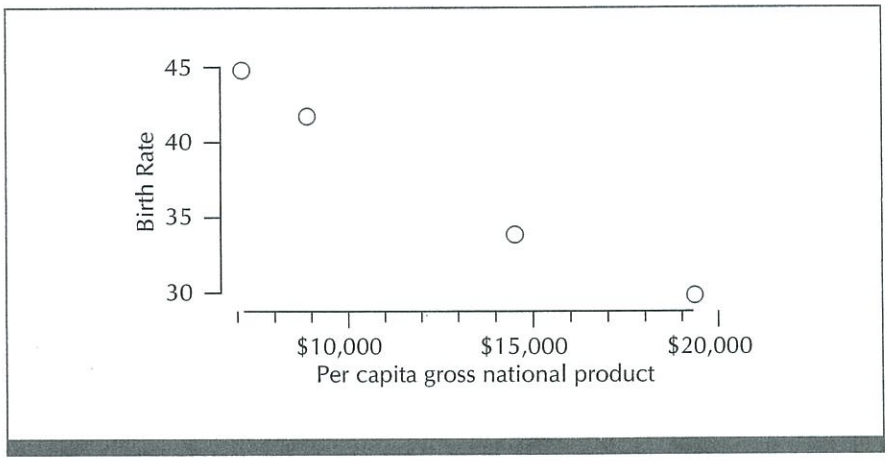

FIGURE 15-1 Scatterplot of birth rate versus per capita gross national product in 1985 for four high-income oil exporters

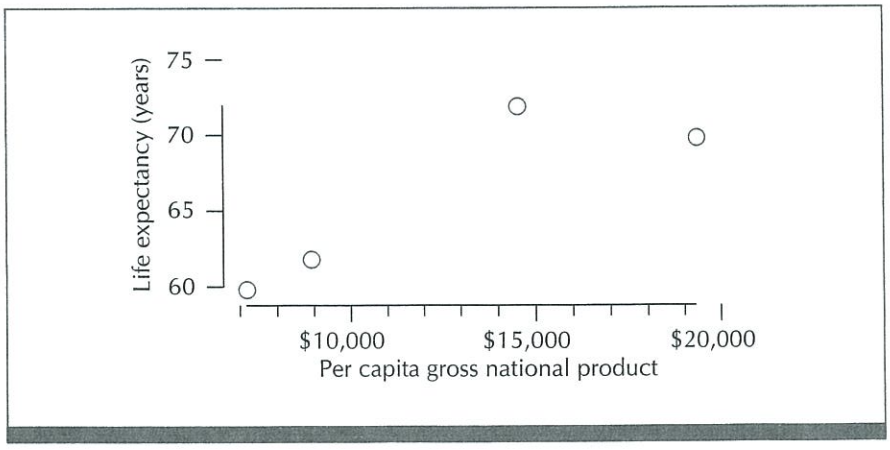

FIGURE 15-2 Scatterplot of life expectancy versus per capita gross national product in 1985 for four high-income oil exporters

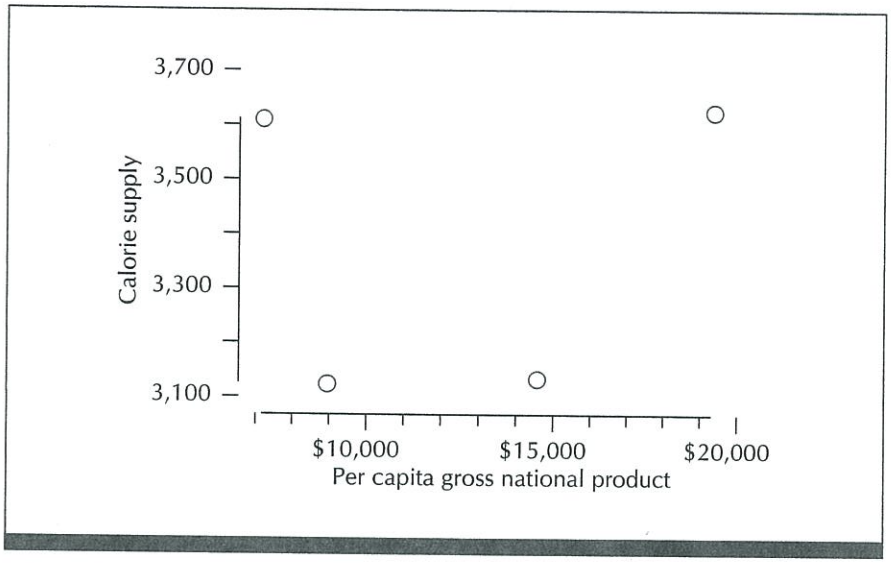

FIGURE 15-3 Scatterplot of daily calorie supply versus gross national product per capita in 1985 for four high-income oil exporters

tion coefficient, then find its value for the three sets of points in Figures 15-1, 15-2, and 15-3.

Suppose we have n pairs of observations (X_i, Y_i), where i goes from 1 to n. We use X_i to denote the ith observation on a variable we call X. Similarly, Y_i denotes the ith observation on a variable we call Y. We want to measure the extent of linear association between the two variables X and Y.

Standardize each value of the first variable by subtracting the sample mean and dividing by the sample standard deviation for that variable. (The mean of a standardized variable equals 0 and the standard deviation equals 1. This is why we call such a variable *standardized.*) Now standardize each value of the second variable. Then for each pair of observations, multiply the standardized values of the two variables. Add up these products, and then divide by the number of pairs minus 1. The result is the *linear correlation coefficient:*

$$\text{Linear correlation coefficient} = \frac{\text{Sum of the products of the two standardized variables}}{\text{Number of pairs} - 1}$$

Let $\bar{X}$ and SD_x denote the sample mean and sample standard deviation, respectively, for the X variable. Similarly, let $\bar{Y}$ and SD_y denote the sample mean and sample standard deviation for the Y variable. Then we can write the formula for the linear correlation coefficient as

$$\text{Linear correlation coefficient} = \frac{\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\text{SD}_x}\right)\left(\frac{Y_i - \bar{Y}}{\text{SD}_y}\right)}{n - 1}$$

Another name for the linear correlation coefficient is Pearson's correlation coefficient. We often refer to it simply as the correlation coefficient, and denote it by r. Alternative calculation formulas for the correlation coefficient are shown below.

The **correlation coefficient,** also called the **linear correlation coefficient** or **Pearson's correlation coefficient,** is a measure of *linear* association between two quantitative variables. If we have n pairs of observations (X_i, Y_i), then we calculate the correlation coefficient r as

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left(\sum_{i=1}^{n} (X_i - \bar{X})^2\right)\left(\sum_{i=1}^{n} (Y_i - \bar{Y})^2\right)}}$$

$$= \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)}}$$

The linear correlation coefficient has no units, and takes values from -1 to 1. A correlation coefficient near 0 suggests there is little or no linear association between the two variables.

A linear correlation coefficient near 1 suggests a strong positive linear association between the two variables. The correlation coefficient equals 1 when and only when all plotted points fall on a straight line with positive slope. The correlation coefficient gives us no information on what this slope is.

A linear correlation coefficient near -1 suggests a strong negative linear association. The correlation coefficient equals -1 when and only when all the points lie on a straight line with negative slope. Again, we cannot determine the slope from the correlation coefficient.

Let's find the correlation coefficient to measure linear association between birth rate and per capita gross national product for the four high-income oil exporters. The calculations are outlined in Table 15-2.

The last two columns in Table 15-2 show the standardized values of gross national product and birth rate. Because of round-off errors in the calculations, the means of our standardized variables may not equal 0 exactly. Similarly, because of rounding errors, these standardized variables may have standard deviations not exactly equal to 1.

At the bottom of Table 15-2, we see that the linear correlation coefficient equals $-.99$. This value indicates a strong negative linear association between birth rate and per capita gross national product for these four countries. As we saw in Figure 15-1, the four plotted points do lie very close to a line with negative slope.

We saw a positive relationship between life expectancy and per capita gross national product in Figure 15-2. The points are not as close to a straight line as the points in Figure 15-1. (United Arab Emirates, with the highest per

TABLE 15-2 Calculating the linear correlation coefficient to measure linear association between per capita gross national product (GNP) and birth rate in 1985 for four high-income, oil-exporting nations

Country	GNP	Birth rate	Standardized GNP	Standardized birth rate
Libya	7,170	45	−.95	1.04
Saudi Arabia	8,850	42	−.65	.61
Kuwait	14,480	34	.37	−.54
United Arab Emirates	19,270	30	1.24	−1.12
Mean	12,442.50	37.75	.0	.0
Standard deviation	5,521.82	6.95	1.0	1.0

Linear correlation coefficient

$$= \frac{(-.95)(1.04) + (-.65)(.61) + (.37)(-.54) + (1.24)(-1.12)}{4 - 1} = -.99$$

capita gross national product, has a life expectancy 2 years shorter than that of Kuwait.) The correlation coefficient equals .88, reflecting the strong positive association between life expectancy and per capita gross national product among these four high-income oil exporters. But .88 is smaller than .99, consistent with our observations that the linear association between life expectancy and per capita gross national product is less than that between birth rate and per capita gross national product among these four countries.

We saw no linear association between daily calorie supply and gross national product per capita in Figure 15-3. The correlation coefficient for these four points is .19. This relatively small value reflects the lack of linear association we saw in the scatterplot for these two variables.

What exactly does the correlation coefficient measure? It measures the extent of clustering of plotted points about a straight line. A correlation coefficient that is large in absolute value suggests strong linear association between the two variables; the variation of points about a line is small relative to the variation in the separate variables. A correlation coefficient near 0 suggests little linear association between the two variables; the variation of points about a line is close to the variation in the separate variables.

Let's discuss these ideas in terms of the scatterplots in Figures 15-4 and 15-5. Figure 15-4 shows a scatterplot of life expectancy versus the logarithm of per capita gross national product for 109 countries. The mean ± 1 standard deviation for life expectancy is graphed near the left vertical axis. The mean ± 1 standard deviation for the logarithm of per capita gross national product is graphed near the top horizontal axis.

A scatterplot of primary school enrollment versus the logarithm of per capita gross national product is shown in Figure 15-5. The mean ± 1 standard deviation is graphed for each variable, as in Figure 15-4.

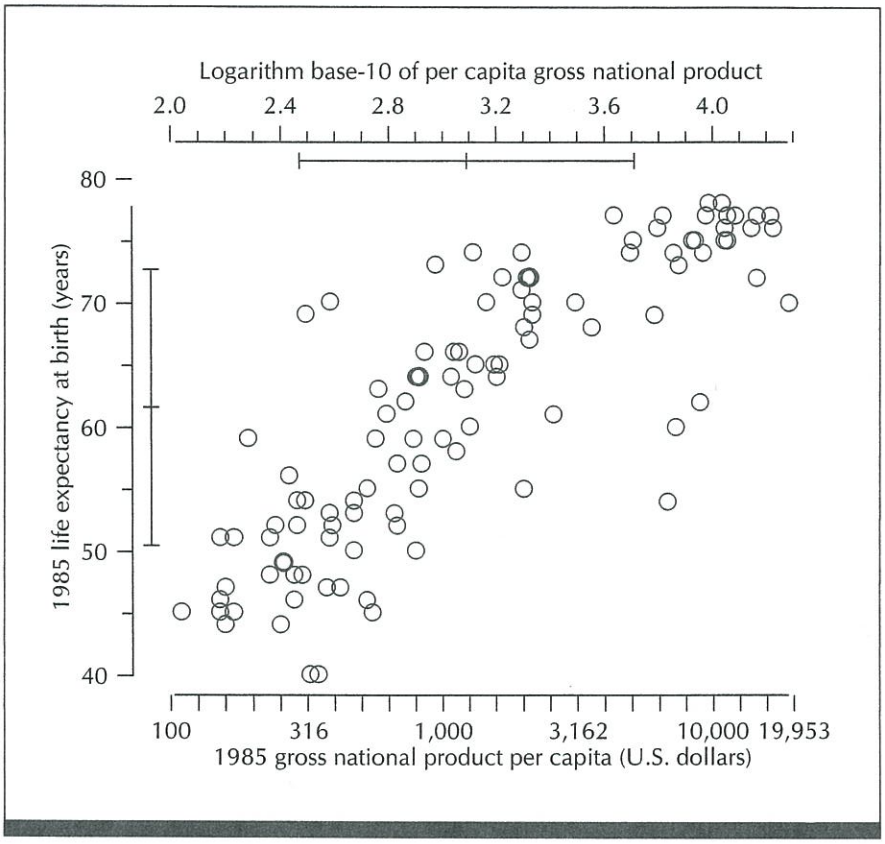

FIGURE 15-4 Scatterplot of life expectancy versus the logarithm of per capita gross national product in 1985 for 109 countries. The mean ± 1 standard deviation for each variable is graphed near the corresponding axis.

We see that there is tighter clustering about a line in Figure 15-4 than in Figure 15-5. The variation about a line drawn through the points in Figure 15-4 is relatively small compared with the variation in the separate variables. In contrast, the variation about a line drawn through the points in Figure 15-5 is close to the variation in the separate variables.

We think there is a stronger linear association illustrated in Figure 15-4 than in Figure 15-5. The correlation coefficients reflect the different impressions we get from these two plots. The correlation coefficient for life expectancy and the logarithm of per capita gross national product is .84. The correlation coefficient for primary school enrollment and the logarithm of per capita gross national product is .49.

Can the correlation coefficient be misleading? Yes, it can. We should always plot two quantitative variables to get a visual feel for their relationship. Then we can use the correlation coefficient to supplement the plot.

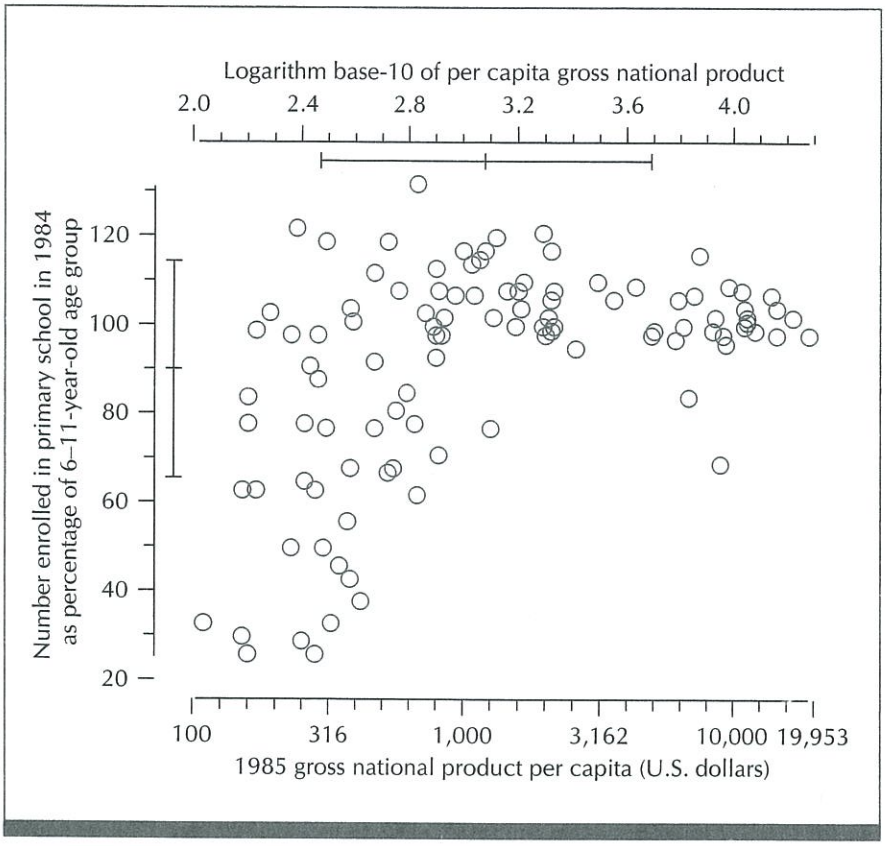

FIGURE 15-5 Scatterplot of number enrolled in primary school in 1984 as percentage of 6–11-year age group and the logarithm of per capita gross national product in 1985 for 104 countries. The mean ± 1 standard deviation for each variable is graphed near the corresponding axis.

Consider the scatterplot of life expectancy versus per capita gross national product in Figure 15-6. The correlation coefficient for the 109 plotted points is .66. By itself, this correlation coefficient might suggest a linear association between these two variables. But we can see in Figure 15-6 a curved relationship. A stronger linear relationship exists between life expectancy and the logarithm of per capita gross national product (Figure 15-4, $r = .84$).

Sometimes a single point or a few points inflate the correlation coefficient (in absolute value) above what it would be if the point(s) were excluded. Consider Figures 15-7 and 15-8, for example. On the vertical axis in each plot is the difference between male and female primary school enrollments in 1985. Overall primary school enrollment is on the horizontal axis.

Figure 15-7 is based on the seven countries in the nonmember economic category with nonmissing information on primary school enrollment. The correlation coefficient for these seven points is .90, a large value. Notice that there

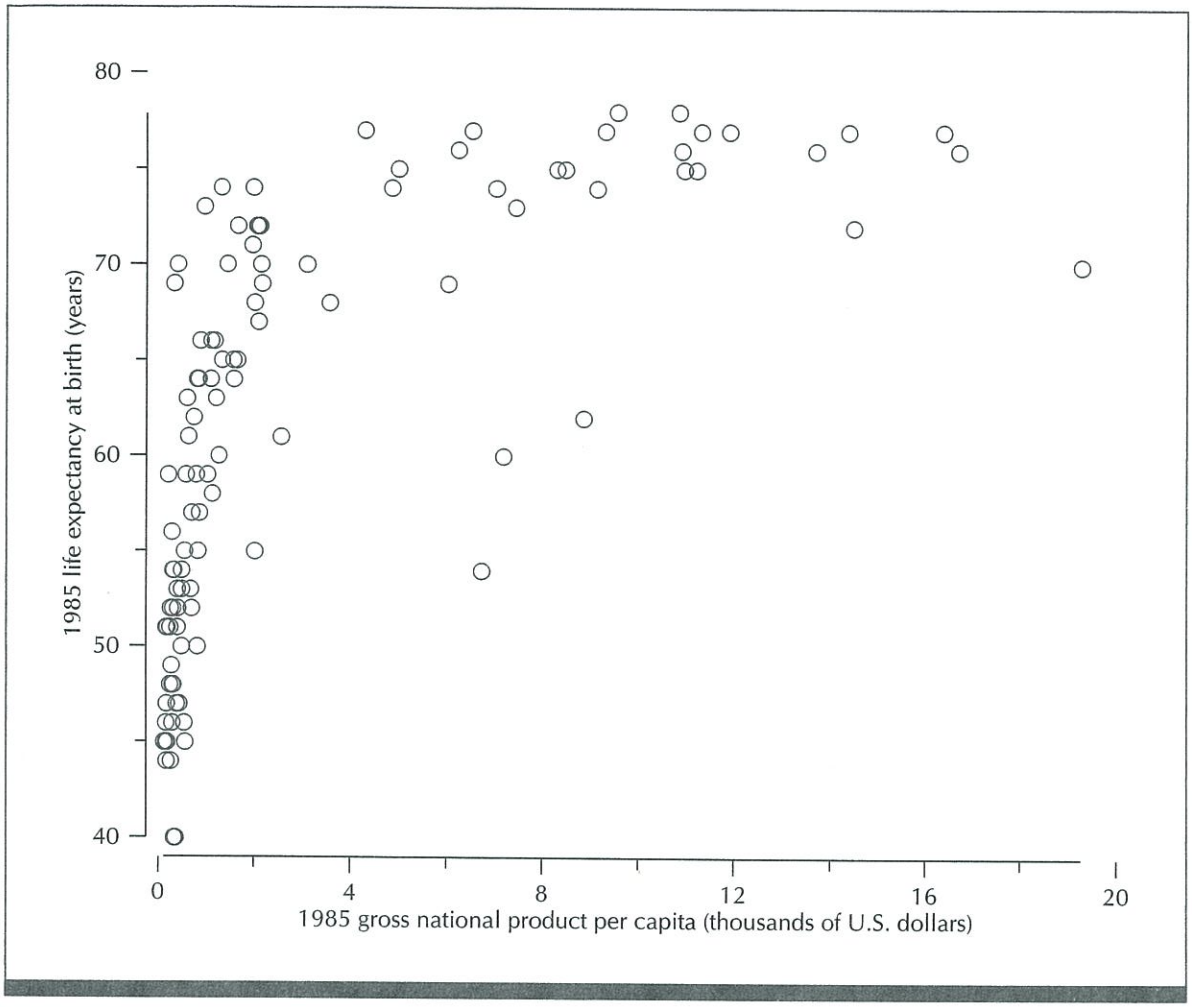

FIGURE 15-6 Scatterplot of life expectancy versus per capita gross national product in 1985 for 109 countries

is a single point by itself in the upper right-hand corner of Figure 15-7. It corresponds to Angola, with a primary school enrollment of 134% and a difference between male and female enrollments of 25%. We might call this point an *outlier:*

An **outlier** is an observation that is far from the other observations.

If we disregard Angola, we get the plot in Figure 15-8. The correlation coefficient for these remaining six points is .40. The large correlation coefficient (.90) for the points in Figure 15-7 results from the relative position of the single point corresponding to Angola. We are unwise to attach much signifi-

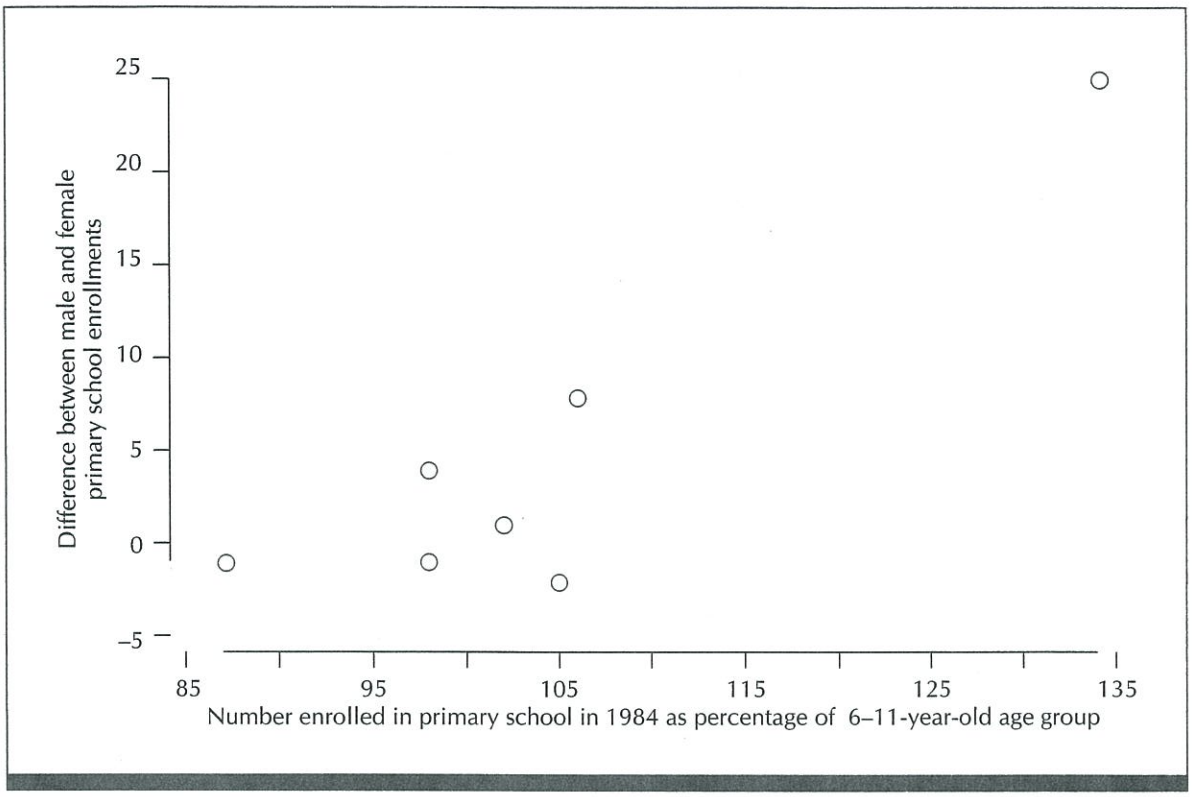

FIGURE 15-7 Scatterplot of the difference between male and female primary school enrollments and overall primary school enrollment in 1984 for seven countries in the nonmember economic category. Two nonmember countries are excluded because of missing values.

cance to a large correlation coefficient that results from the position of a single point.

One point or a few points can also pull a correlation coefficient closer to 0 than it would be if the point(s) were excluded. Figure 15-9 shows a scatterplot of infant mortality rate and per capita gross national product in 1985 for the 20 upper-middle-income countries with nonmissing values for both variables. The correlation coefficient is $-.05$, about as close to 0 as we might expect to see.

Examining Figure 15-9, we see a general trend of decreasing infant mortality rates with increasing per capita gross national product. There is a single striking exception—the point in the upper right-hand corner. This exception is Oman, with per capita gross national product of $6,730 and an infant mortality rate of 109. Removing Oman, we get the plot in Figure 15-10. The correlation coefficient for the remaining 19 points is $-.47$. This is more consistent with our impression from the plot.

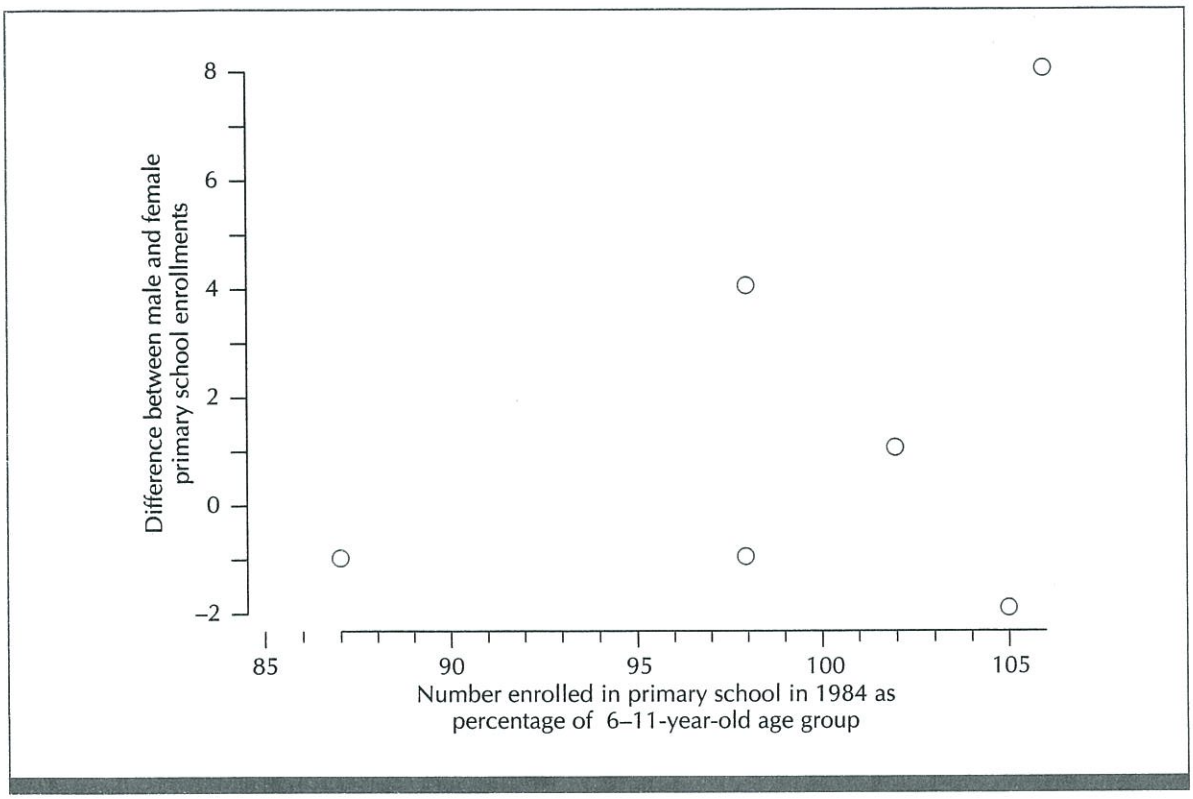

FIGURE 15-8 This scatterplot is the same as the one in Figure 15-7 except that the outlying point in the upper right-hand corner of Figure 15-7 has been excluded.

We can learn a lot in the process of finding out why an outlier is an outlier. Oman has a relatively high per capita gross national product, but a high infant mortality rate typical of the low-income countries. The high infant mortality rate makes Oman different from the other upper-middle-income countries. In fact, in the 1985 *World Development Report,* the World Bank classified Oman as a high-income oil exporter (World Bank, 1985). Recall from Part I that the high-income oil exporters are similar to the low-income countries for some indicators. Therefore, in discussing the relationship between infant mortality and gross national product for upper-middle-income countries, we might want to consider Oman as a special case.

We have to be careful with outliers. We should not exclude a case from an analysis just because it is different from the others. By judicious exclusion of cases, we may "see" characteristics in our data set that are not really there. This is, of course, *not* the purpose of data analysis.

Let's look now at how the difference between female and male life expectancy varies with overall life expectancy. A scatterplot of the difference be-

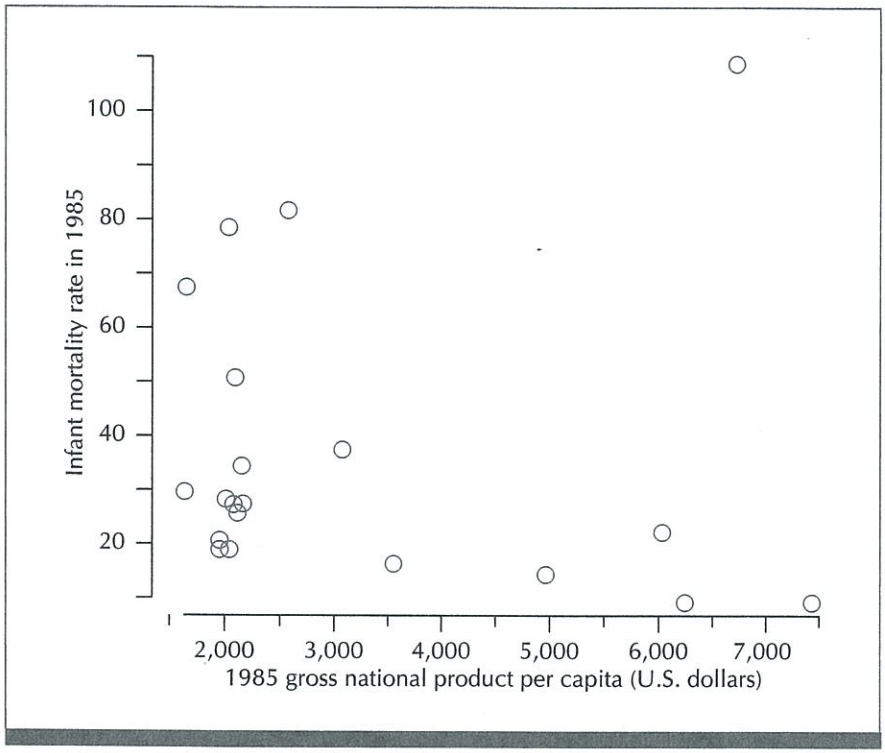

FIGURE 15-9 Scatterplot of infant mortality rate and per capita gross national product in 1985 for 20 upper-middle-income countries. The range of values for each variable is indicated by the line on the corresponding axis.

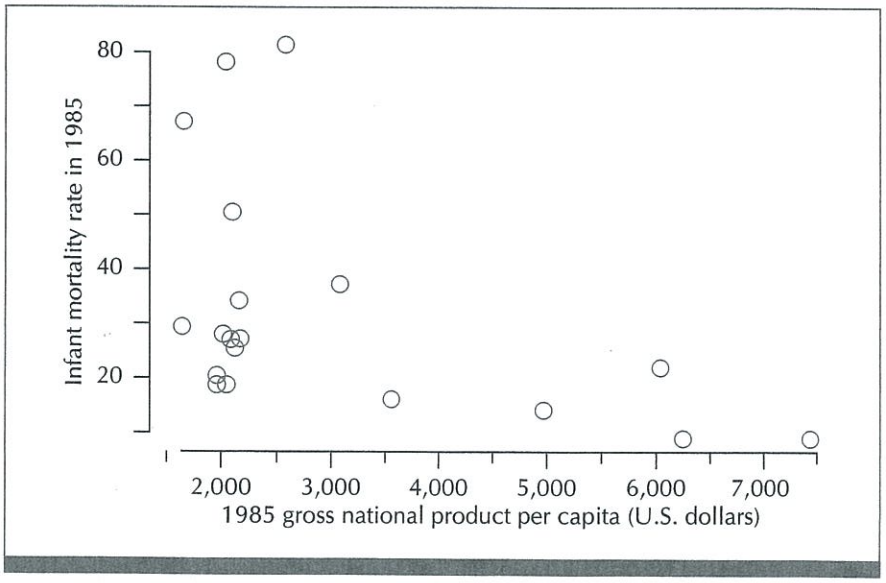

FIGURE 15-10 This scatterplot is the same as the one in Figure 15-9 except that the outlying point in the upper right-hand corner of Figure 15-9 is excluded.

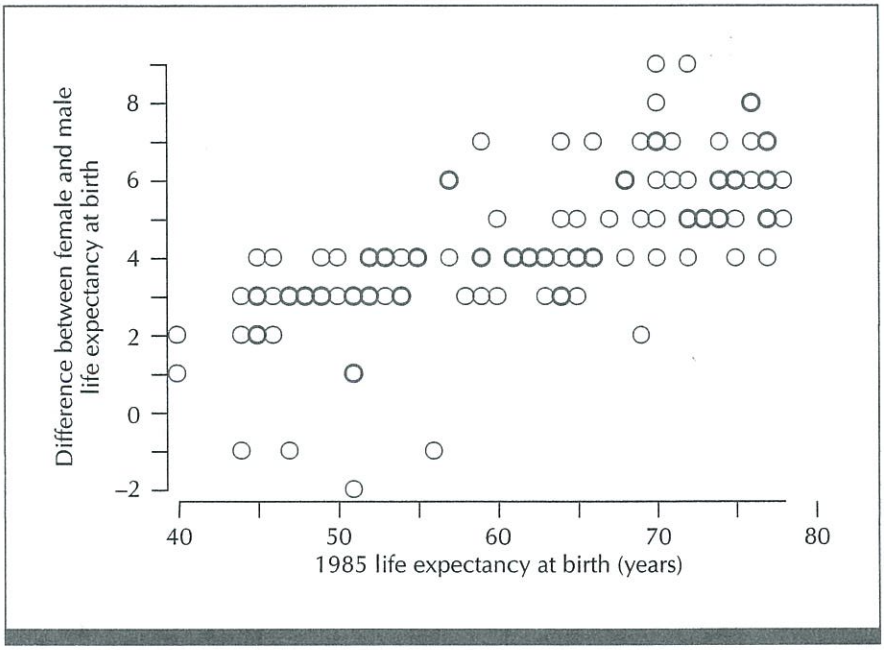

FIGURE 15-11 Scatterplot of the difference between female and male life expectancies and overall life expectancy at birth in 1985 for 125 countries

tween female and male life expectancies and overall life expectancy is shown in Figure 15-11 for 125 countries. There is a fairly strong increasing relationship, with a correlation coefficient of .70.

A scatterplot of the difference between female and male life expectancies and overall life expectancy is shown in Figure 15-12 for the 19 industrial market countries. We no longer see an increasing linear relationship; the correlation coefficient is $-.05$, very close to 0.

There is a suggestion of another type of relationship between the two variables in Figure 15-12. We see that the differences between female and male life expectancies are largest for countries with overall life expectancy of 76 years (for instance, United States life expectancy was 80 years for females, 72 years for males). The differences are smaller for overall life expectancies shorter than 76 years, as well as for overall life expectancies longer than 76 years. This plot gives us a suggestion of a *quadratic* relationship. Such a relationship is not indicated at all by the correlation coefficient, which measures only *linear* association. (Figure 15-12 might lead us to hope that as life expectancies continue to lengthen, the gap in expected life span between females and males will decrease. This is speculation, of course, because the data plotted in Figure 15-12 are not sufficient for drawing any such conclusion.)

In this section we have discussed the correlation coefficient as a descriptive statistic: a measure of linear association between two quantitative variables.

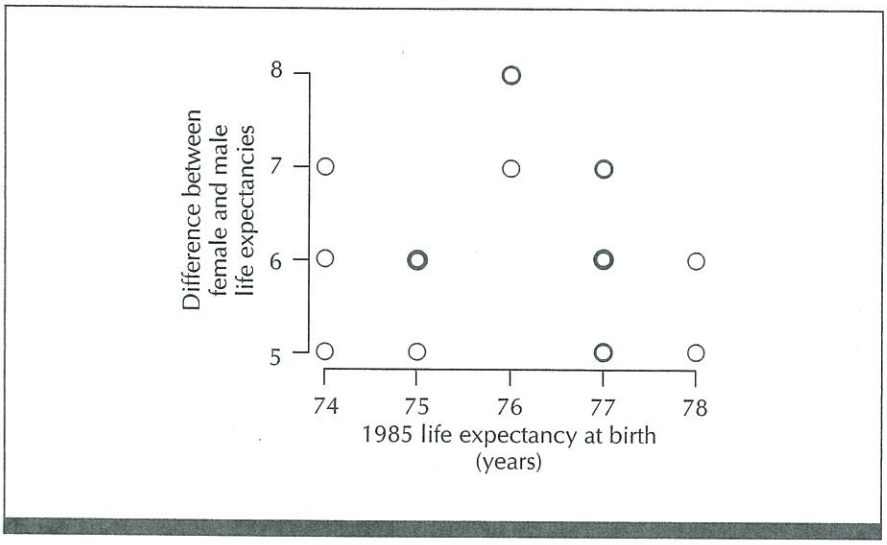

FIGURE 15-12 Scatterplot of the difference between female and male life expectancies and overall life expectancy in 1985 for 19 industrial market countries

When certain assumptions about the variables and the sampling process are met, we can test the null hypothesis that the linear correlation coefficient for two variables is 0. (We used World Bank data in this section to illustrate the correlation coefficient as a descriptive statistic measuring the extent of linear association between two variables. We cannot test hypotheses using the World Bank data set. We have information on the entire population of World Bank countries, rather than on a randomly selected sample of a population.) In Section 15-2, we discuss a parametric test that a linear correlation coefficient equals 0. We can apply this test when our sample meets assumptions described in that section.

15-2

A Parametric Test That a Linear Correlation Coefficient Equals Zero

Suppose we have n independent pairs of observations (X_i, Y_i). We let X_i denote the ith observation on a variable we call X. Y_i denotes the ith observation on a variable Y. We want to test the null hypothesis that the linear correlation coefficient between observations on X and Y is 0.

Let's assume that X_1 through X_n represent a random sample from a Gaussian distribution, and Y_1 through Y_n a random sample from another Gaussian distribution. [We must also assume that the pairs (X_i, Y_i) represent a random sample from what we call a bivariate normal, or bivariate Gaussian, distribution. See, for example, Brownlee (1965, Chapter 12).] With these assumptions,

If $6 < D < 64$, say the results are consistent with the null hypothesis that thickness and stiffness are independent in this fabric.
If $D \leq 6$ or $D \geq 64$, say the results are inconsistent with the null hypothesis, suggesting that thickness and stiffness are not independent in this fabric.

From Table 15-3 we see that $D = 8$, which is in the acceptance region. The p-value $= 2P(D \leq 8) = .102$. Based on this p-value, we might say the results are borderline, especially since the sample size is small. Certainly Figure 15-15 suggests positive association between thickness and stiffness in this flame-retardant fabric.

For the sake of illustration, suppose we try the large-sample test in Example 15-2. We calculate the test statistic:

$$\text{Test statistic} = \frac{8 - \dfrac{6^3 - 6}{6}}{\sqrt{\dfrac{6^2(6 + 1)^2(6 - 1)}{36}}} = -1.72$$

Looking at Table B for the standard Gaussian distribution, we see our approximate p-value is .0854, somewhat smaller than the p-value of .102 we get with the exact distribution of D under the null hypothesis.

In Section 15-4, we discuss the method of least squares for fitting a straight line to a sample of pairs of observations. We also discuss parametric hypothesis testing for the straight-line model.

15-4

Simple Linear Regression and the Method of Least Squares

In many situations we want not only to measure the extent of linear association between two variables, but also to estimate the linear relationship between them. We would like to model one variable as a straight-line function of another, using the *method of least squares.*

Suppose we have n pairs of observations (X_i, Y_i) on two variables. We plot the observations and a linear association seems reasonable. Imagine drawing a straight line $Y = b_0 + b_1X$ through the cloud of plotted points. Here, b_0 denotes the intercept and b_1 the slope of the line. We will use the method of least squares to determine b_0 and b_1.

For any given value X_i of the first variable, we can use the straight-line model to predict the associated Y value. Let $\hat{Y}_i$ denote this predicted or estimated mean Y value. Then $\hat{Y}_i = b_0 + b_1X_i$. The difference $Y_i - \hat{Y}_i$ is a residual, measuring how far the estimated mean Y value is from the actual Y value for the ith observation.

> A **residual** is the difference between a Y value and a predicted or estimated mean Y value, when a variable Y is modeled as a function of one or more other variables.

Using the method of least squares, we find the values of b_0 and b_1 that minimize the sum of the squared residual differences $(Y_i - \hat{Y}_i)^2$.

Suppose we have n pairs of observations (X_i, Y_i). Using the **method of least squares** to fit a straight line $Y = b_0 + b_1X$, we find constants b_0 and b_1 to minimize

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (b_0 + b_1X_i))^2$$

Provided our observations include at least two distinct values of the X variable, we can find the least squares intercept b_0 and slope b_1. We can calculate these least squares values of b_0 and b_1 using the formulas

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \quad \text{and} \quad b_0 = \bar{Y} - b_1\bar{X}$$

Let's illustrate the method of least squares with an example.

EXAMPLE 15-3 For each of ten streets with bike lanes, investigators measured the distance between the center line and a cyclist in the bike lane. They used photography to determine the distance between a cyclist and a passing car on those same ten streets, recording all distances in feet. The results are shown below (Devore, 1982, pages 432–433; from "Effects of Bike Lanes on Driver and Bicyclist Behavior," *ASCE Transportation Eng. J.*, 1977, pages 243–256).

Street	1	2	3	4	5	6	7	8	9	10
Center line to cyclist (feet)	12.8	12.9	12.9	13.6	14.5	14.6	15.1	17.5	19.5	20.8
Car to cyclist (feet)	5.5	6.2	6.3	7.0	7.8	8.3	7.1	10.0	10.8	11.0

A plot of the observations is shown in Figure 15-16. Based on a visual inspection of this scatterplot, a linear relationship between the two variables seems reasonable.

Let the X variable be the distance from the center line to the cyclist. Let the Y variable be the distance from the car to the cyclist. We will use the method of least squares to model Y as a straight-line function of X. The necessary calculations are outlined in Table 15-4.

We see from the bottom of Table 15-4 that the least squares line is $Y = -2.18 + .66X$. The positive slope of .66 agrees with the positive association between the two variables that we see in Figure 15-16. The intercept of -2.18 has no physical meaning in this example: We cannot let the distance X from the center line to a cyclist in the bike lane be 0 feet, because the distance Y from the car to the cyclist cannot be -2.18 feet!

The least squares line for our example is plotted in Figure 15-17. The vertical distances from the points (X_i, Y_i) to the least squares line are indicated by dashed lines. These distances are the values of the residuals $Y_i - \hat{Y}_i$. The

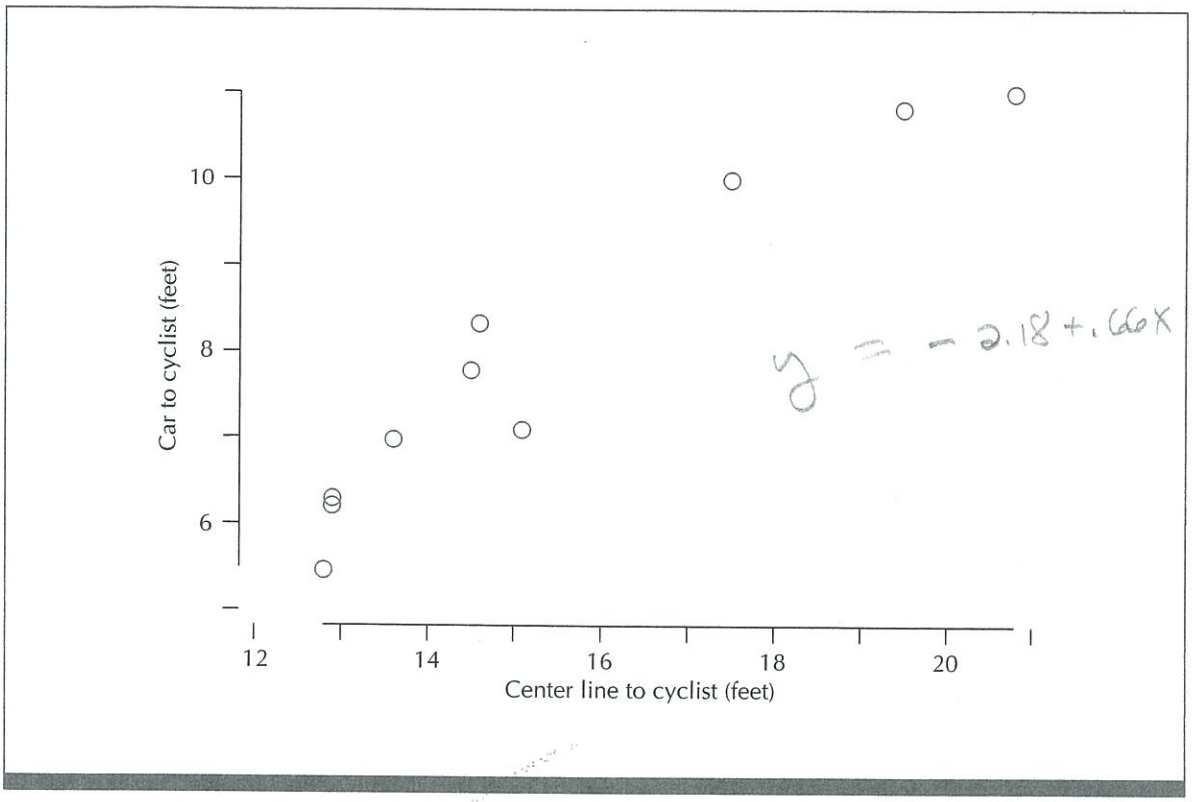

FIGURE 15-16 Scatterplot of the distance from car to cyclist and the distance from center line to cyclist in Example 15-3

TABLE 15-4 Calculations for finding the least squares intercept b_0 and slope b_1 for Example 15-3

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
12.8	5.5	−2.62	−2.5	6.550	6.8644
12.9	6.2	−2.52	−1.8	4.536	6.3504
12.9	6.3	−2.52	−1.7	4.284	6.3504
13.6	7.0	−1.82	−1.0	1.820	3.3124
14.5	7.8	−.92	−.2	.184	.8464
14.6	8.3	−.82	.3	−.246	.6724
15.1	7.1	−.32	−.9	.288	.1024
17.5	10.0	2.08	2.0	4.160	4.3264
19.5	10.8	4.08	2.8	11.424	16.6464
20.8	11.0	5.38	3.0	16.140	28.9444
$\bar{X} = 15.42$	$\bar{Y} = 8$		Total:	49.14	74.416

$b_1 = \dfrac{49.14}{74.416} = .66$ $\quad b_0 = 8 - (.66)(15.42) = -2.18$

Least squares line: $Y = -2.18 + .66X$

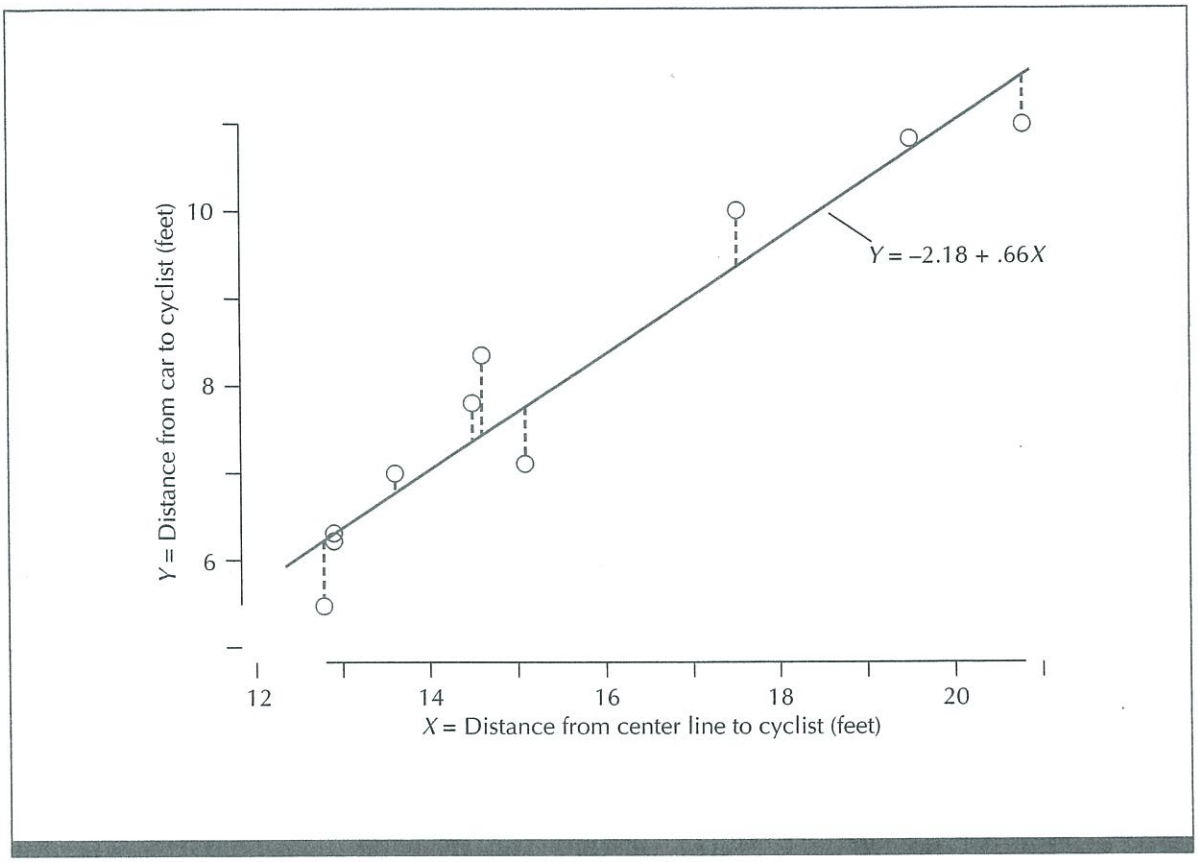

FIGURE 15-17 Scatterplot of distance from car to cyclist versus distance from center line to cyclist in Example 15-3. Also shown is the least squares line.

least squares line is best in the sense of minimizing the sum of the squares of these vertical distances, or residuals.

The slope b_1 in the equation $Y = b_0 + b_1X$ has units equal to the units of the Y variable, divided by the units of the X variable. The slope represents the change in the Y variable for each unit increase in the X variable.

The units of the intercept b_0 are the units of the Y variable. The intercept has a physical interpretation only if there are values of the X variable very close to 0 and if it is possible for the X variable to equal 0.

The method of least squares requires no assumptions. We need to make assumptions about the observations only if we want to test hypotheses about the straight-line model. Suppose we do want to test hypotheses about the straight-line model. We use the term *simple linear regression* to refer to the process of fitting a straight-line model by the method of least squares and testing hypotheses about the model.

Simple linear regression refers to fitting a straight-line model by the method of least squares and then assessing the model.

The *classical assumptions for simple linear regression* are these: Suppose we have n pairs of observations (X_i, Y_i). We observe values of the X variable with no error. Y_1 through Y_n are independent random variables, Y_i coming from a Gaussian distribution with mean equal to $\beta_0 + \beta_1 X_i$ and variance equal to σ^2.

The parameters (or unknown numbers) β_0 and β_1 are the intercept and slope, respectively, of the line describing the relationship between X and Y. The variance σ^2 describes the random variation of the Y values about that line.

We want to estimate the intercept β_0 and the slope β_1, as well as the variance σ^2. We also want to test the null hypothesis that $\beta_1 = 0$ and the null hypothesis that $\beta_0 = 0$.

We estimate β_0 and β_1 using the least squares estimates b_0 and b_1 given previously. We estimate σ^2 with the residual mean square s_r^2:

$$s_r^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

where $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$ is the ith residual, the difference between the observed and estimated Y values.

To test the hypotheses $H_0: \beta_1 = 0$ and $H_a: \beta_1 \neq 0$, we use the test statistic

$$\text{Test statistic(1)} = \frac{b_1}{\sqrt{\dfrac{s_r^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}}$$

Under the null hypothesis that $\beta_1 = 0$, test statistic(1) has the t distribution with $n - 2$ degrees of freedom. Values of test statistic(1) far from 0 (in either the positive or negative direction) are inconsistent with the null hypothesis that the slope β_1 equals 0. Note that when we ask whether the slope β_1 equals 0, we implicitly assume that the intercept β_0 is in the model.

To test the hypotheses $H_0: \beta_0 = 0$ and $H_a: \beta_0 \neq 0$, we use the test statistic

$$\text{Test statistic(0)} = \frac{b_0}{\sqrt{s_r^2\left(\dfrac{1}{n} + \dfrac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}}$$

Under the null hypothesis that $\beta_0 = 0$, test statistic(0) has the t distribution with $n - 2$ degrees of freedom. Extreme values of test statistic(0), far from 0 in either the positive or negative direction, are inconsistent with the null hypothesis that the intercept β_0 equals 0. When we ask whether the intercept β_0 equals 0, we implicitly assume that the slope β_1 is in the model.

EXAMPLE 15-3 *(continued)*

Let's illustrate these ideas by continuing with Example 15-3. Some calculations we need are summarized in Table 15-5.

Before testing hypotheses, we should check our model assumptions. The straight-line relationship between X (distance from center line to cyclist) and

TABLE 15-5 Calculations for simple linear regression in Example 15-3

X	Y	$\hat{Y} = -2.18 + .66X$	$e = Y - \hat{Y}$	$e^2 = (Y - \hat{Y})^2$
12.8	5.5	6.27	−.77	.5929
12.9	6.2	6.33	−.13	.0169
12.9	6.3	6.33	−.03	.0009
13.6	7.0	6.80	.20	.0400
14.5	7.8	7.39	.41	.1681
14.6	8.3	7.46	.84	.7056
15.1	7.1	7.79	−.69	.4761
17.5	10.0	9.37	.63	.3969
19.5	10.8	10.69	.11	.0121
20.8	11.0	11.55	−.55	.3025
			Total:	2.7120

$s_r^2 = \dfrac{2.7120}{10 - 2} = .339$ Degrees of freedom = 10 − 2 = 8

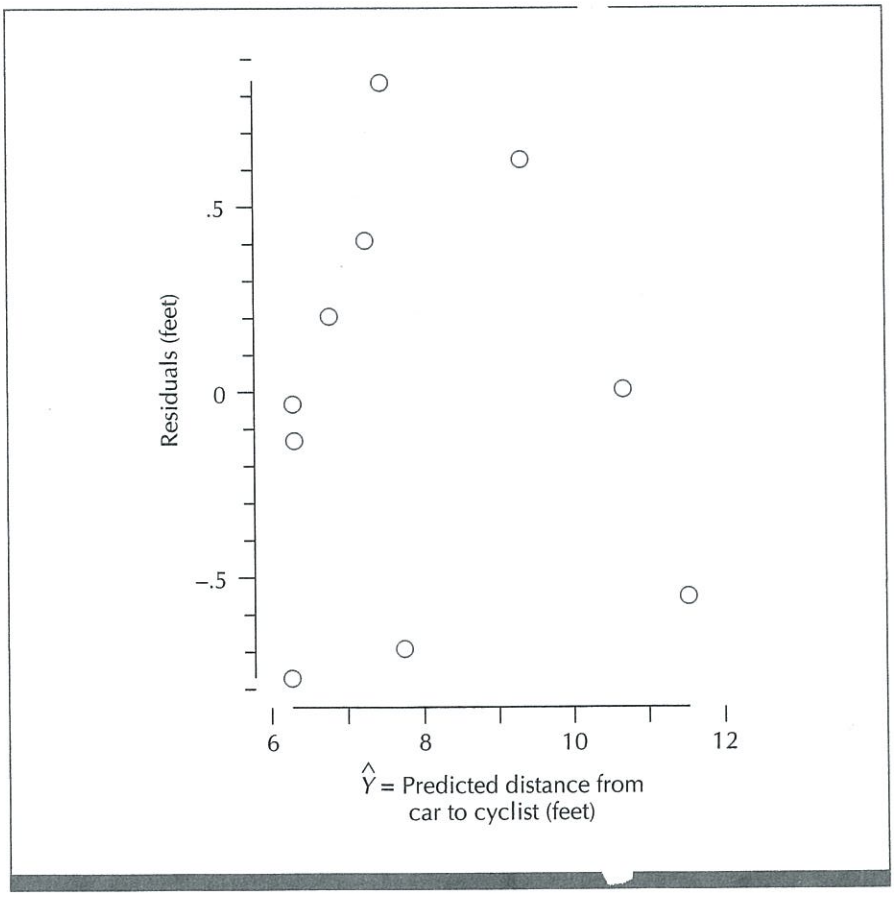

FIGURE 15-18 Scatterplot of residuals versus predicted Y values in Example 15-3

Y (distance from car to cyclist) seems reasonable from Figure 15-17. As another check, we can look at a plot of residuals versus predicted Y values, as in Figure 15-18. If the straight-line model with constant variation holds, the residuals should represent random variation or noise. A residual plot showing a pattern that does not look like random variation or noise suggests that the straight-line model may not be appropriate. We cannot see any particular pattern in Figure 15-18, so this residual plot gives us no reason to doubt the straight-line model.

We must assume that the X variable is measured without error; it seems reasonable that the investigators could measure the distance from the center line to the cyclist with minimal error. Figures 15-17 and 15-18 give us no basis to doubt the assumption that each Y_i has the same variance σ^2.

Also, we assume that the Y_i's are independent; we cannot assess this independence assumption without more information on how the experiment was conducted. What suggestions would you have for carrying out this experiment, in order to ensure independence and reduce the effects of extraneous sources of variation?

We must assume, in addition, that Y_i comes from a Gaussian distribution with mean $\beta_0 + \beta_1 X_i$ and variance σ^2 or, equivalently, that $Y_i - (\beta_0 + \beta_1 X_i)$ comes from a Gaussian distribution with mean 0 and variance σ^2. We use the residual $e_i = Y_i - \hat{Y}_i$ to estimate $Y_i - (\beta_0 + \beta_1 X_i)$. A dot plot of the residuals is shown in Figure 15-19. From this figure, we see no reason to doubt the Gaussian assumption.

Let's test the null hypothesis that the slope β_1 equals 0. Using calculations outlined in Tables 15-4 and 15-5, we see that

$$\text{Test statistic}(1) = \frac{.66}{\sqrt{\frac{.339}{74.416}}} = 9.8$$

Referring to Table C for the t distribution with 8 degrees of freedom, we see that our p-value is less than .01. The results are inconsistent with the null hypothesis that the slope β_1 equals 0.

Now let's test the null hypothesis that the intercept β_0 equals 0. We see that

$$\text{Test statistic}(0) = \frac{-2.18}{\sqrt{.339\left(\frac{1}{10} + \frac{(15.42)^2}{74.416}\right)}} = -2.1$$

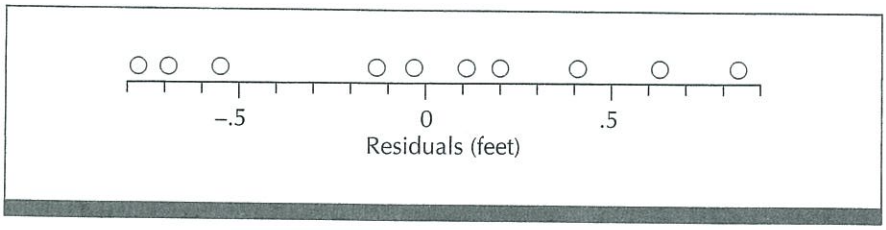

FIGURE 15-19 Dot plot of the residuals in Example 15-3

Again referring to Table C for the t distribution with 8 degrees of freedom, we see that our p-value is between .05 and .10, which is borderline. We will suppose that β_0 is not 0, that we do need a nonzero intercept in our model.

We usually calculate a statistic called R^2 in simple linear regression. R^2 is the square of the simple linear correlation coefficient r for the X and Y variables, so $R^2 = r^2$. R^2 is the proportion of the variation in the Y variable explained by the straight-line model.

> In simple linear regression, the square of the linear correlation coefficient, denoted $\boldsymbol{R^2}$, is the proportion of the variation in the response variable accounted for, or explained, by the straight-line model.

In Example 15-3, the correlation coefficient r for the two variables is .96. Therefore, $R^2 = (.96)^2 = .92$. We say about 92% of the variation in distances between cars and cyclists is explained by the linear relationship between that variable and the distance from center line to cyclist. This is a fairly large value of R^2. From our analysis, including the scatterplot of the data values and the residual plots, it seems that a straight-line model is very reasonable in Example 15-3.

In Section 15-5, we discuss the relation between correlation and simple linear regression.

15-5 Correlation and Simple Linear Regression

Suppose once again that we have n pairs of observations (X_i, Y_i). The linear correlation coefficient r measures the extent of linear association between the X and Y variables. It measures how closely the plotted points cluster about a line. We call this line the *standard deviation line* (Freedman, Pisani, and Purves, 1978, page 122).

The standard deviation line is the line most of us would draw freehand through a cloud of plotted points. It passes through the point $(\bar{X}, \bar{Y})$ corresponding to the sample means for the two variables. The slope of the standard deviation line is

$$\text{Slope of standard deviation line} = \begin{cases} \dfrac{SD_y}{SD_x} & \text{if the association is positive} \\ -\dfrac{SD_y}{SD_x} & \text{if the association is negative} \end{cases}$$

where SD_y and SD_x denote the sample standard deviations of the Y and X variables, respectively.

Figure 15-20 shows a plot of the observations in Example 15-1. (Recall that investigators used two methods to determine the amount of breast milk ingested by each of 14 babies.) The standard deviation line is also shown. This line goes through the point $(\bar{X}, \bar{Y}) = (1{,}616.4, 1{,}449.1)$. The slope of the stan-

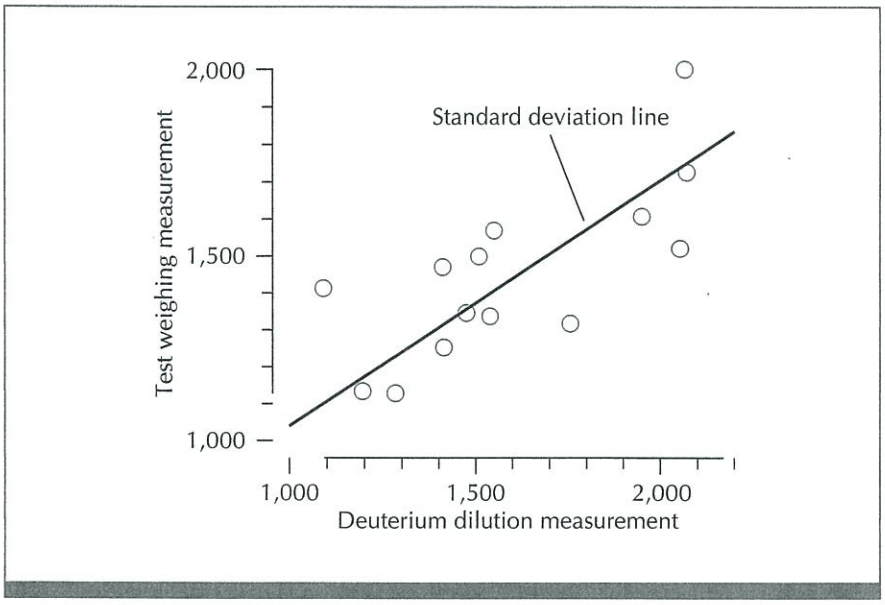

FIGURE 15-20 Plot of measurement of milk ingested using the test weighing method versus measurement using the deuterium dilution technique. The standard deviation line is also shown.

dard deviation line is positive because the association between the two variables is positive:

$$\text{Slope of standard deviation line in Example 15-1: } \frac{\text{SD}_y}{\text{SD}_x} = \frac{234}{353} = .66$$

How does the standard deviation line compare with the least squares line? We can show that another (equivalent) formula for the slope of the least squares line is

$$\text{Slope of the least squares line} = b_1 = r\frac{\text{SD}_y}{\text{SD}_x}$$

The slope of the least squares line equals the absolute value of the linear correlation coefficient r times the slope of the standard deviation line. Since the absolute value of r is in the range from 0 to 1, we see that the least squares line is less steep than the standard deviation line.

In Example 15-1, we found the linear correlation coefficient to be $r = .77$. Therefore, the slope of the least squares line is

$$\text{Slope of least squares line in Example 15-1} = (.77)(.66) = .51$$

The least squares line and the standard deviation line are both plotted in Figure 15-21. Note that both lines pass through the point $(\bar{X}, \bar{Y})$. The least squares line is indeed less steep than the standard deviation line. For a large value of X (greater than $\bar{X}$), the corresponding value of Y predicted by the least squares

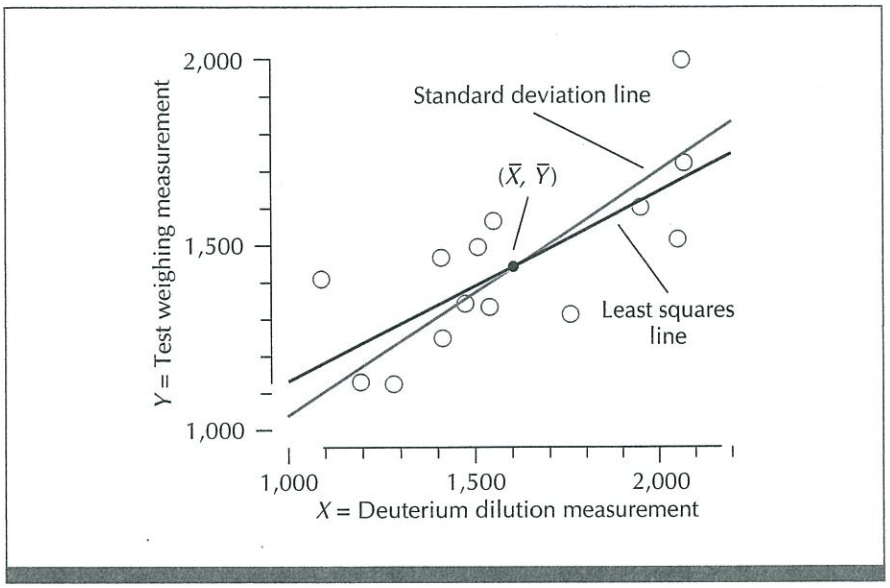

FIGURE 15-21 Scatterplot of the observations in Example 15-1. The least squares line and the standard deviation line are shown.

line is less than that predicted by the standard deviation line. For a small value of X (less than $\bar{X}$), the value of Y predicted by the least squares line is greater than that predicted by the standard deviation line.

Let's look at another example.

EXAMPLE 15-4

Foresters recorded two characteristics of 20 stands of pine trees (Myers, 1986, page 68; from Burkhart et al., 1972). One characteristic was the number of pine trees per acre; the other was the average diameter of pine trees 4.5 feet above the ground (units not given). The values of the two variables are listed in Example 15-5, in Section 15-6. Some descriptive statistics are shown in Table 15-6.

A plot of the observations is given in Figure 15-22. We see that there is a negative relationship between number of pine trees per acre and the average diameter of the trees. The standard deviation line and the least squares line are also shown in the figure, the standard deviation line steeper than the least squares line. For a value of X (number of pine trees per acre) greater than $\bar{X}$, the corresponding value of Y (average diameter) predicted by the least squares line is greater than that predicted by the standard deviation line. For a value of X less than $\bar{X}$, the value of Y predicted by the least squares line is less than that predicted by the standard deviation line.

We use the least squares line to estimate the average value of the Y variable corresponding to a particular value of the X variable. This estimated Y value is generally less extreme than what we might expect by a freehand sketch of a line through the plotted points, because our freehand sketches tend

TABLE 15-6 Descriptive statistics for Example 15-4

X = Number of pine trees per acre		Y = Average diameter 4.5 feet above the ground		
$\bar{X} = 671.5$	$SD_x = 136.9$	$\bar{Y} = 6.265$	$SD_y = .739$	$r = -.252$

Slope of standard deviation line $= -\dfrac{.739}{136.9} = -.00540$

Slope of least squares line $= (-.252)(.00540) = -.00136$

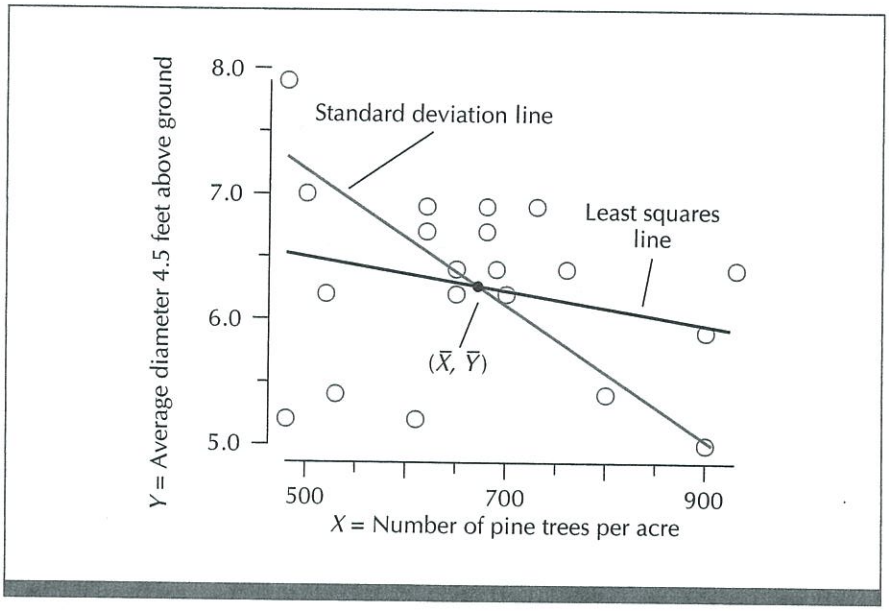

FIGURE 15-22 Scatterplot of the observations in Example 15-4. The least squares line and the standard deviation line are shown.

to be closer to the standard deviation line. The geneticist Sir Francis Galton (1822–1911) noticed this when he studied the sizes of seeds and their offspring and when he studied the heights of fathers and sons. Extremely tall fathers, for instance, had sons who were shorter than they, on average; extremely short fathers had sons who were taller than they, on average. We could turn this around and say extremely tall sons had fathers who were shorter than they, on average; extremely short sons had fathers who were taller than they, on average. Galton called this "regression towards mediocrity" or *regression toward the mean.* The term is unfortunately ambiguous; all it means is that the least squares line is less steep than the standard deviation line, as we have noted. It is from Galton that we get the term *regression,* as in *simple linear regression.* (For a discussion of regression toward the mean or the regression fallacy, see Freedman, Pisani, and Purves, 1978, pages 158–162.)

A final comment on the relationship between correlation and simple linear regression: The test of the null hypothesis that the slope is 0 in simple

linear regression (Section 15-4) is equivalent to the test that the linear correlation coefficient is 0 (Section 15-2).

In Section 15-6, we present a very brief introduction to multiple regression.

15-6 A Brief Introduction to Multiple Regression

Suppose we record values of variables X_1 through X_k and Y for each of n observations. We might denote the values for the ith observation by $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i)$. We want to model the Y variable as a function of the variables X_1 through X_k, say:

$$Y = b_0 + b_1X_1 + \cdots + b_kX_k$$

We call this a *linear model;* the model is linear in the constants b_0, b_1 through b_k.

> A **linear model** is a model that is linear in the parameters, the unknown constants in the model.

We call the process of fitting and assessing such a model *multiple regression.*

> By **multiple regression** we mean the process of modeling a quantitative variable as a function of several other variables, and assessing the model. We consider only linear models.

Using such a model, we can estimate the value of the Y variable for any set of values of X_1 through X_k. We use the notation $\hat{Y}_i$ to denote the estimated, or predicted, value of Y for the ith observation:

$$\hat{Y}_i = b_0 + b_1X_{1i} + \cdots + b_kX_{ki}$$

We let $e_i = Y_i - \hat{Y}_i$ denote the ith residual, the difference between Y and the estimated or predicted value of Y for the ith observation.

Using the *method of least squares* we find the values of the constants b_0 through b_k that minimize the sum of the squared residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (b_0 + b_1X_{1i} + \cdots + b_kX_{ki}))^2$$

We can find unique values of b_0, b_1 through b_k to minimize the sum of squared residuals, as long as there are at least $k + 1$ distinct sets of values of X_1 through X_k.

We do not need to make any assumptions to fit a linear model using the method of least squares. However, if we want to test hypotheses about the model, we do need to make some assumptions.

The *classical assumptions for multiple regression* are these: The variables X_1 through X_k are measured without error. Y_1 through Y_n represent independent observations from Gaussian distributions, Y_i from a Gaussian distribution

with mean $\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$ and variance σ^2. We use the method of least squares to find the least squares estimates b_0, b_1 through b_k of β_0, β_1 through β_k, respectively.

If these assumptions are met, we can test the hypotheses H_0: $\beta_j = 0$ and H_a: $\beta_j \neq 0$, where β_j is one of the parameters in the model. When we ask whether the parameter β_j equals 0, we implicitly assume that the other parameters are in the model. The test statistic is

$$\text{Test statistic}(j) = \frac{b_j}{\text{SE}(b_j)}$$

where b_j is the least squares estimate of β_j and $\text{SE}(b_j)$ is the standard error, or estimated standard deviation, of b_j. If the null hypothesis is true, test statistic(j) has the t distribution with $n - p$ degrees of freedom, where p is the number of parameters (unknown β_i's) in the model. (We have $p = k + 1$ in the model above.)

The calculations for multiple linear regression are so extensive that we use a computer to perform them.

Let's consider an example.

EXAMPLE 15-5

Foresters studied 20 stands of pine trees. For each stand, they recorded the age of the stand (units not given), the average height in feet of dominant trees, the number of pine trees per acre, and the average diameter 4.5 feet above the ground (units not given). The results are shown below (Myers, 1986, page 68; from Burkhart et al., 1972).

Stand	Age	Height	Number	Average diameter
1	19	51.5	500	7.0
2	14	41.3	900	5.0
3	11	36.7	650	6.2
4	13	32.2	480	5.2
5	13	39.0	520	6.2
6	12	29.8	610	5.2
7	18	51.2	700	6.2
8	14	46.8	760	6.4
9	20	61.8	930	6.4
10	17	55.8	690	6.4
11	13	37.3	800	5.4
12	21	54.2	650	6.4
13	11	32.5	530	5.4
14	19	56.3	680	6.7
15	17	52.8	620	6.7
16	15	47.0	900	5.9
17	16	53.0	620	6.9
18	16	50.3	730	6.9
19	14	50.5	680	6.9
20	22	57.7	480	7.9

The foresters wanted to model Y (average diameter) as a function of one or more of the variables X_1 (age), X_2 (height), X_3 (number), X_4 (age × number), and X_5 (height/number).

Table 15-7 shows the linear correlation coefficient for each pair of the six variables Y, X_1, X_2, X_3, X_4, and X_5. We see that the largest correlation coefficient (in absolute value) for average diameter Y with any other variable is the correlation coefficient of .840 for Y and X_5 = height/number.

We will go through the steps of a *backward regression analysis* for this problem. We start with a model that includes all the variables that we think might be important for estimating Y. If any parameter (other than the intercept β_0) seems to be 0, we will drop from the model the one with the largest p-value. Then we will fit a new model excluding the dropped parameter and its corresponding variable. We continue dropping one variable at a time and refitting until all the parameters (other than β_0) appear to be different from 0.

First let's consider the model

$$\text{Expected value of } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Results of a multiple regression analysis using the Student Edition of Minitab are shown in Table 15-8.

TABLE 15-7 The linear correlation coefficient for each pair of variables in Example 15-5

	Y Average diameter	X_1 Age	X_2 Height	X_3 Number	X_4 Age × Number	X_5 Height/ Number
Y	1.000					
X_1	.675	1.000				
X_2	.773	.876	1.000			
X_3	−.252	.016	.229	1.000		
X_4	.244	.678	.755	.732	1.000	
X_5	.840	.735	.634	−.579	.056	1.000

TABLE 15-8 Results of the first multiple regression analysis for Example 15-5

Parameter	Variable	Parameter estimate	Standard error	Test statistic	p-value
β_1	Age	.0526	.1683	.31	.759
β_2	Height	.08246	.04035	2.04	.060
β_3	Number	.003224	.002532	1.27	.224
β_4	Age × Number	−.0002817	.0002300	−1.22	.241
β_5	Height/Number	16.03	27.89	.57	.575
β_0	Intercept	1.233	1.619	.76	.459

p = Number of parameters = 6 n = Sample size = 20

Degrees of freedom for tests of hypotheses = $n - p$ = 14

TABLE 15-9 Results of the second multiple regression analysis for Example 15-5

Parameter	Variable	Parameter estimate	Standard error	Test statistic	*p*-value
β_2	Height	.07438	.03004	2.48	.026
β_3	Number	.002795	.002065	1.35	.196
β_4	Age × Number	−.0002131	.00006746	−3.16	.006
β_5	Height/Number	22.93	16.53	1.39	.186
β_0	Intercept	1.507	1.321	1.14	.272

p = Number of parameters = 5 n = Sample size = 20

Degrees of freedom for tests of hypotheses = $n - p = 15$

We see that the least squares estimated model is

$$\hat{Y} = 1.23 + .0526X_1 + .0825X_2 + .00322X_3 - .000282X_4 + 16.0X_5$$

where the parameter estimates are shown to three significant figures. The test statistic for a parameter tests the null hypothesis that the parameter is 0, when all the other parameters are in the model. The largest p-value, .759, is consistent with the null hypothesis that β_1 equals 0. That is, if all the other variables are in the model, it looks like we do not need to include X_1 (age).

We drop X_1 and consider the model:

$$\text{Expected value of } Y = \beta_0 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5$$

Results of the multiple regression analysis are shown in Table 15-9.

The least squares estimated model is

$$\hat{Y} = 1.51 + .0744X_2 + .00280X_3 - .000213X_4 + 22.9X_5$$

The largest p-value (ignoring the intercept) is .196, consistent with the null hypothesis that β_3 equals 0. That is, if the other parameters in Table 15-9 are included in the model, it looks like we need not include X_3 (number of pine trees per acre).

We drop X_3 and consider the model

$$\text{Expected value of } Y = \beta_0 + \beta_2X_2 + \beta_4X_4 + \beta_5X_5$$

Results of the multiple regression analysis are given in Table 15-10.

The least squares estimated model is

$$\hat{Y} = 3.24 + .0974X_2 - .000169X_4 + 3.47X_5$$

The largest p-value, .684, is consistent with the null hypothesis that β_5 equals 0. If the other parameters in Table 15-10 are included in the model, it looks like we do not need to include X_5 (height/number). We drop X_5, even though it had the highest linear correlation coefficient with Y in Table 15-7.

For our final analysis, we drop X_5 and consider the model

$$\text{Expected value of } Y = \beta_0 + \beta_2X_2 + \beta_4X_4$$

Table 15-11 shows the results of the multiple regression analysis.

TABLE 15-10 Results of the third multiple regression analysis for Example 15-5

Parameter	Variable	Parameter estimate	Standard error	Test statistic	*p*-value
β_2	Height	.09741	.02540	3.84	.001
β_4	Age × Number	−.0001689	.00006052	−2.79	.013
β_5	Height/Number	3.467	8.374	.41	.684
β_0	Intercept	3.2357	.3467	9.33	.000

p = Number of parameters = 4 n = Sample size = 20

Degrees of freedom for tests of hypotheses = $n - p = 16$

TABLE 15-11 Results of the final multiple regression analysis for Example 15-5

Parameter	Variable	Parameter estimate	Standard error	Test statistic	*p*-value
β_2	Height	.10691	.01058	10.11	.000
β_4	Age × Number	−.00018975	.00003256	−5.83	.000
β_0	Intercept	3.2605	.3330	9.79	.000

p = Number of parameters = 3 n = Sample size = 20

Degrees of freedom for tests of hypotheses = $n - p = 17$

The least squares estimated model is

$$\hat{Y} = 3.26 + .107X_2 - .000190X_4$$

All the *p*-values in Table 15-11 are less than .001. This suggests that all three parameters—β_0, β_2, and β_4—are necessary to the model.

A descriptive statistic that we often use in multiple regression is the *multiple regression coefficient* or *coefficient of determination,* denoted R^2. The multiple regression coefficient has an interpretation similar to that of R^2 in simple linear regression. The multiple regression coefficient R^2 is the proportion of the variation in the Y values that is explained by the multiple regression model.

> The **multiple regression coefficient** or **coefficient of determination, R^2,** is the proportion of the variation in the response variable that is accounted for, or explained, by the multiple regression model.

For the final model in Table 15-11, we have $R^2 = .87$, a fairly large value. About 87% of the variation in average diameters can be explained by the model that includes height of the dominant trees and age times the number of pine trees per acre.

The predicted Y values and residuals for this final model are shown in Table 15-12. A plot of residuals versus predicted Y values is shown in Figure 15-23. This scatterplot looks like "noise" because we cannot see any pattern or relationship between the residuals and the predicted Y values. This residual

TABLE 15-12 List of the values of Y, the predicted Y values, and the residuals from the model $\hat{Y} = 3.26 + .107X_2 - .000190X_4$ in Example 15-5

Y = Average diameter	$\hat{Y}$ = Predicted value of Y	$e = Y - \hat{Y}$ = Residual
7.0	6.96	.04
5.0	5.29	−.29
6.2	5.83	.37
5.2	5.52	−.32
6.2	6.15	.05
5.2	5.06	.14
6.2	6.34	−.14
6.4	6.25	.15
6.4	6.34	.06
6.4	7.00	−.60
5.4	5.27	.13
6.4	6.47	−.07
5.4	5.63	−.23
6.7	6.83	−.13
6.7	6.91	−.21
5.9	5.72	.18
6.9	7.04	−.14
6.9	6.42	.48
6.9	6.85	.05
7.9	7.43	.47

plot gives us no reason to doubt the equal-variance assumption or the adequacy of the model.

A dot plot of the residuals from this final model is shown in Figure 15-24. The residuals have a fairly symmetrical distribution concentrated around zero. This plot gives us no reason to doubt the Gaussian assumption.

We must assume that height, age, and number of pine trees per acre are observed without error, and that the observed values of average diameter are independent. We would need additional information about the experiment to assess these assumptions. What suggestions would you make about experimental design to help meet model assumptions and reduce extraneous sources of variation?

A scatterplot matrix of the variables height (X_2), age × number (X_4), and average diameter (Y) is shown in Figure 15-25. We can see the positive association between average diameter and height (correlation coefficient = .77). However, there is only a weak association between average diameter and age × number (correlation coefficient = .24). It is common in multiple regression situations that simple bivariate plots do not give us a good impression of which variables are needed in the model.

Notice the strong positive association between height and age × number (correlation coefficient = .75). Sometimes predictor variables are even more

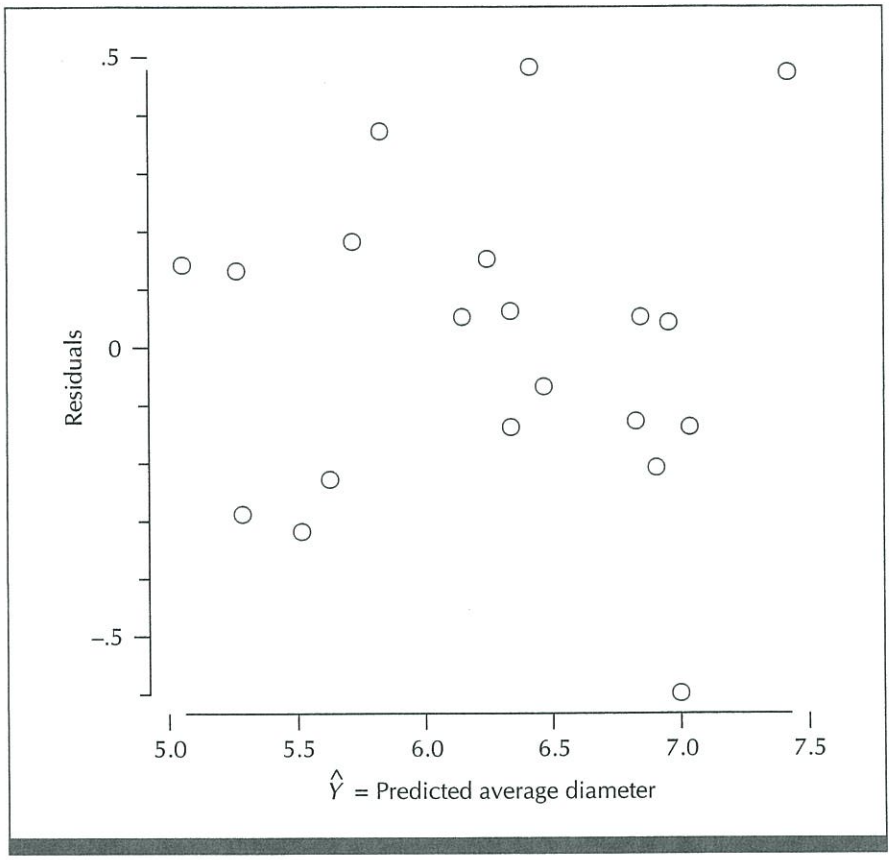

FIGURE 15-23 Plot of residuals versus predicted Y values for the model $\hat{Y} = 3.26 + .107X_2 - .000190X_4$ in Example 15-5

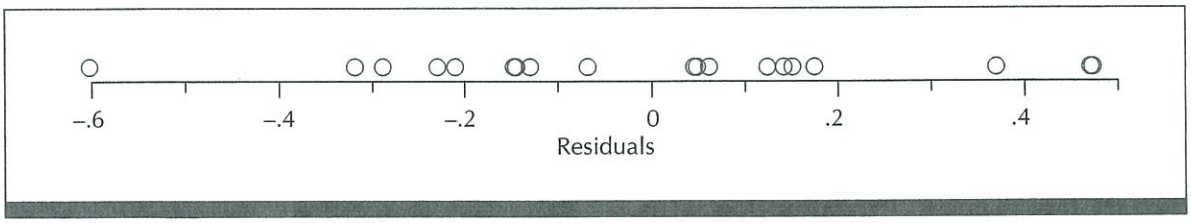

FIGURE 15-24 Dot plot of residuals from the model $\hat{Y} = 3.26 + .107X_2 - .000190X_4$ in Example 15-5

highly correlated. If we try to include in a model two or more predictor variables that are highly correlated with one another, we can run into trouble. It is a problem we call *multicollinearity*.

> When two or more predictor variables in a multiple regression analysis are highly correlated with one another, we can get errors in the analysis, a problem called **multicollinearity.**

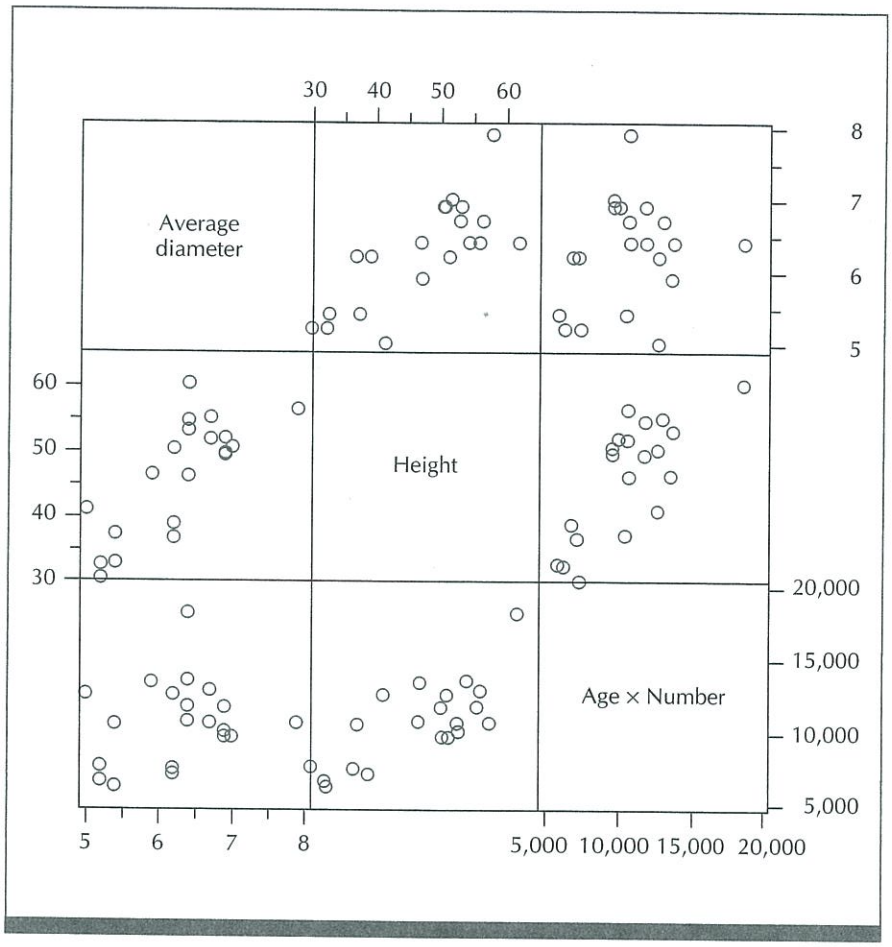

FIGURE 15-25 Scatterplot matrix of X_2 = height, X_4 = age × number, and Y = average diameter in Example 15-5

There are many aspects of multiple regression that we have not addressed, in addition to possible multicollinearity. We have not discussed how to choose predictor variables, how to evaluate observations that especially influence the fitted model, or how to deal with outliers, for instance. Multiple regression requires a course to itself. This section is meant just to give an idea of what it is about. For more information see, for example, Draper and Smith (1981), Daniel and Wood (1980), Myers (1986), or Rawlings (1988).

Summary of Chapter 15

The linear correlation coefficient is a measure of linear association between two quantitative variables. The correlation coefficient is a descriptive statistic.

It may be used to measure linear association between two variables, even in situations when formal statistical inference may not be appropriate.

When certain assumptions about the observations and sampling process are satisfied, we can make inferences about the linear association between two variables. A parametric test that the linear correlation coefficient equals 0 is also a test that the two variables are independent.

The rank correlation coefficient is a measure of association between two quantitative variables, based on ranks. A nonparametric test of independence between two quantitative variables may be based on the rank correlation coefficient.

The method of least squares is one way to model one quantitative variable as a straight-line function of another. We can use the method of least squares to fit a straight line without making any assumptions about the observations. Hypothesis testing in simple linear regression does require that we make assumptions about the observations and sampling process.

Considering the relation between correlation and simple linear regression, we introduce the standard deviation line and compare it with the least squares line. Examples illustrate the idea of regression toward the mean.

In multiple regression, we model a quantitative variable as a function of several other variables. We consider only linear models—that is, models that are linear in the parameters or unknown constants. The method of least squares allows us to calculate a multiple regression model for a set of observations. Hypothesis testing in multiple regression requires that we make assumptions about the observations and sampling process.

```
* NOTE *  age*num is highly correlated with other  predictor variables

The regression equation is
avediam = 1.23 + 0.053 age + 0.0825 height + 0.00322 number -0.000282 age*num
          + 16.0 ht/num

Predictor       Coef       Stdev    t-ratio        p
Constant       1.233       1.619       0.76    0.459
age           0.0526      0.1683       0.31    0.759
height       0.08246     0.04035       2.04    0.060
number      0.003224    0.002532       1.27    0.224
age*num   -0.0002817   0.0002300      -1.22    0.241
ht/num         16.03       27.89       0.57    0.575

s = 0.2952      R-sq = 88.2%     R-sq(adj) = 84.1%

Analysis of Variance

SOURCE        DF          SS          MS         F        p
Regression     5      9.1651      1.8330     21.03    0.000
Error         14      1.2204      0.0872
Total         19     10.3855


CONTINUE? y
SOURCE        DF      SEQ SS
age            1      4.7388
height         1      1.4684
number         1      2.3984
age*num        1      0.5307
ht/num         1      0.0288

Unusual Observations
Obs.      age   avediam       Fit Stdev.Fit  Residual   St.Resid
  2      14.0    5.0000    5.4633    0.1899   -0.4633     -2.05R
 10      17.0    6.4000    6.9457    0.1254   -0.5457     -2.04R

R denotes an obs. with a large st. resid.
```

FIGURE M15-3 Output for the first multiple regression model specified in Example 15-5

For the command

```
MTB>  regress 'avediam' 5 'age' 'height' &
CONT> 'number' 'age*num' 'ht/num' c10 c11;
SUBC> residuals C12.
```

Minitab will produce the output in Figure M15-3, and save standardized residuals in column 10, predicted values in column 11, and residuals in column 12. We can use these saved values in plots to check model assumptions.

Exercises for Chapter 15

For all exercises, plot the observations in any ways that seem reasonable. Describe the population(s) sampled, whether real or hypothetical. State the assumptions for each test of hypotheses. Do these assumptions seem reasonable? What additional information would you like to have about the experiment? Describe the results of your analysis.

EXERCISE 15-1

In Exercise 4-12, we looked at sodium content and potassium content (no units given) in perspiration of ten healthy women (Oja and Nyblom, 1989; from Johnson and Wichern, 1982, page 182):

Woman	1	2	3	4	5	6	7	8	9	10
Sodium	48.5	65.1	47.2	53.2	55.5	36.1	24.8	33.1	47.4	54.1
Potassium	9.3	8.0	10.9	12.2	9.7	7.9	14.0	7.6	8.5	11.3

a. Plot the observations.

b. Calculate the linear correlation coefficient for the sodium and potassium measurements. Test the null hypothesis that the linear correlation coefficient between sodium and potassium levels in perspiration of healthy women is 0.

c. Calculate the rank correlation coefficient for the sodium and potassium measurements. Carry out a nonparametric test that sodium and potassium levels in perspiration of healthy women are independent.

d. Compare your answers to parts (b) and (c). Discuss your findings.

EXERCISE 15-2

In Exercise 4-11, we considered carbon monoxide concentration (parts per million) and benzo(a)pyrene concentration (μg per 1,000 cubic meters) in 16 different air samples from Herald Square in New York City (Devore, 1982, page 457; from "Carcinogenic Air Pollutants in Relation to Automobile Traffic in New York City," *Environmental Science and Technology,* 1971, pages 145–150). The results are shown below as pairs of readings for each air sample: (carbon monoxide, benzo(a)pyrene).

(2.8, .5)	(15.5, .1)	(19.0, .8)	(6.8, .9)	(5.5, 1.0)
(5.6, 1.1)	(9.6, 3.9)	(13.3, 4.0)	(5.5, 1.3)	(12.0, 5.7)
(5.6, 1.5)	(19.5, 6.0)	(11.0, 7.3)	(12.8, 8.1)	(5.5, 2.2)
(10.5, 9.5)				

a. Plot the observations.

b. Calculate the linear correlation coefficient for the two substances. Test the null hypothesis that the linear correlation coefficient between carbon monoxide readings and benzo(a)pyrene readings at Herald Square under similar conditions is 0.

c. Calculate the rank correlation coefficient for the two substances. Test the null hypothesis that carbon monoxide readings and benzo(a)pyrene readings at Herald Square under similar conditions are independent.

d. Compare your answers to parts (b) and (c). Discuss your findings.

EXERCISE 15-3

Scientists wanted to compare the drop net catch method and the sweep net catch method of collecting grasshoppers (Walpole and Myers, 1989, page 439; from the Department of Entomology, Virginia Polytechnic Institute and State

University). They recorded the average number of grasshoppers caught in each of 17 field quadrants using the two methods. They also recorded the average height of plants in each quadrant. All measurements were made the same day.

Quadrant	Average drop net catch	Average sweep net catch	Average height of plants (centimeters)
1	18.00	4.15	52.7
2	8.88	2.02	42.1
3	2.00	.16	34.8
4	20.00	2.33	27.6
5	2.38	.26	45.9
6	2.75	.57	97.5
7	3.33	.70	102.1
8	1.00	.14	97.8
9	1.33	.12	88.3
10	1.75	.11	58.7
11	4.13	.56	42.4
12	12.88	2.45	31.3
13	5.38	.45	31.8
14	28.00	6.69	35.4
15	4.75	.87	64.5
16	1.75	.15	25.2
17	.13	.02	36.4

a. Construct a scatterplot matrix of these three variables.

b. Calculate the linear correlation coefficient for each pair of variables.

c. Calculate the rank correlation coefficient for each pair of variables.

d. Compare your answers to parts (b) and (c). Discuss your findings.

EXERCISE 15-4 Body weight and heart weight are shown below for each of 19 normal woodchucks (Walpole and Myers, 1989, page 398; from the Department of Veterinary Medicine and the Statistics Consulting Center, Virginia Polytechnic Institute and State University).

Woodchuck	Body (grams)	Heart (grams)	Woodchuck	Body (grams)	Heart (grams)
1	4,050	11.2	11	3,690	10.8
2	2,465	12.4	12	2,800	14.2
3	3,120	10.5	13	2,775	12.2
4	5,700	13.2	14	2,170	10.0
5	2,595	9.8	15	2,370	12.3
6	3,640	11.0	16	2,055	12.5
7	2,050	10.8	17	2,025	11.8
8	4,235	10.4	18	2,645	16.0
9	2,935	12.2	19	2,675	13.8
10	4,975	11.2			

a. Plot the observations.

b. Calculate the linear correlation coefficient for body weight and heart weight. Test the null hypothesis that the linear correlation coefficient for body weight and heart weight in normal woodchucks is 0.

c. Calculate the rank correlation coefficient for body weight and heart weight. Carry out a nonparametric test of the null hypothesis that body weight and heart weight are independent in normal woodchucks.

d. Compare your answers to parts (b) and (c). Discuss your results.

EXERCISE 15-5

Investigators simultaneously measured wind speed (m/s) on the ground and via Seasat satellite at each of 12 times (Milton and Arnold, 1986, pages 325–326; from "Mapping Ocean Winds by Radar," *NASA Tech Briefs,* Fall 1982, page 27).

Ground measurement:	4.46	3.99	3.73	3.29	4.82	6.71	4.61	3.87	3.17	4.42	3.76	3.30
Satellite measurement:	4.08	3.94	5.00	5.20	3.92	6.21	5.95	3.07	4.76	3.25	4.89	4.80

a. Plot the observations.

b. Calculate the linear correlation coefficient for the ground and satellite measurements. Test the null hypothesis that the linear correlation coefficient for ground and satellite measurements of wind speed is 0.

c. Calculate the rank correlation coefficient for the ground and satellite measurements. Test the null hypothesis that ground and satellite measurements of wind speed are independent.

d. Compare your answers to parts (b) and (c). Discuss your findings.

EXERCISE 15-6

Researchers measured inulin clearance (ml/min) of seven living kidney donors and the recipients of their kidneys (Hollander and Wolfe, 1973, page 239; from Shelp et al., 1970).

Recipient:	61.4	63.3	63.7	80.0	77.3	84.0	105.0
Donor:	70.8	89.2	65.8	67.1	87.3	85.1	88.1

a. Plot the observations in any ways that seem helpful.

b. Calculate the linear correlation coefficient for recipient and donor inulin clearance. Test the null hypothesis that the linear correlation coefficient for recipient and donor inulin clearance is 0.

c. Calculate the rank correlation coefficient for recipient and donor inulin clear-

ance. Carry out a nonparametric test that recipient and donor inulin clearance measurements are independent.

d. Compare your answers to parts (b) and (c) and discuss your findings.

EXERCISE 15-7 Investigators wanted to study the effects of illumination on a person's ability to perform a task (Devore, 1982, page 300; from "Performance of Complex Tasks Under Different Levels of Illumination," *J. Illuminating Eng.,* 1976, pages 235–242). A volunteer inserted a fine-tipped probe into the eyehole of a needle, ten times with low light and a black background and ten times with more light and a white background. The average time (units not given) at each light level is shown below for each of nine volunteers.

Volunteer:	1	2	3	4	5	6	7	8	9
Higher light level:	25.85	28.84	32.05	25.74	20.89	41.05	25.01	24.96	27.47
Lower light level:	18.23	20.84	22.96	19.68	19.50	24.98	16.61	16.07	24.59

a. Plot the observations in any ways that seem helpful.

b. Calculate the linear correlation coefficient for the two sets of times. Test the null hypothesis that the linear correlation coefficient of average times under the two light levels is 0.

c. Calculate the rank correlation coefficient for the two sets of times. Carry out a nonparametric test of the null hypothesis that the average times under the two light levels are independent.

d. Compare your answers to parts (b) and (c) and discuss your findings.

EXERCISE 15-8 Consider the measurements of thickness and stiffness on six samples of a flame-retardant fabric in Example 15-2, plotted in Figure 15-15.

a. Calculate the linear correlation coefficient for these two variables. Compare the linear correlation coefficient with the rank correlation coefficient calculated in Example 15-2.

b. Find the least squares line modeling stiffness as a function of thickness.

c. Test the null hypothesis that the slope in the straight-line model is 0.

d. Test the null hypothesis that the intercept in the straight-line model is 0.

e. What percentage of the variation in stiffness is explained by the straight-line model? (That is, what is R^2?)

f. Discuss your findings.

EXERCISE 15-9 In a study of the operation of a factory, investigators recorded 25 observations of amount of steam used per month and average atmospheric temperature (Draper and Smith, 1981, page 9):

Steam used (pounds)	Average tempera-ture (°F)	Steam used (pounds)	Average tempera-ture (°F)
10.98	35.3	9.57	39.1
11.13	29.7	10.94	46.8
12.51	30.8	9.58	48.5
8.40	58.8	10.09	59.3
9.27	61.4	8.11	70.0
8.73	71.3	6.83	70.0
6.36	74.4	8.88	74.5
8.50	76.7	7.68	72.1
7.82	70.7	8.47	58.1
9.14	57.5	8.86	44.6
8.24	46.4	10.36	33.4
12.19	28.9	11.08	28.6
11.88	28.1		

a. Plot steam versus temperature.

b. Calculate the linear correlation coefficient r for steam and temperature.

c. Use the method of least squares to model steam used as a straight-line function of average temperature.

d. Use residual plots to assess the fit of the model.

e. Discuss your findings.

EXERCISE 15-10 Exercise 4-8 described an experiment studying plastic spools used in electric motors. Wire is wound around the spools. When current passes through the wire, the temperature of the spool rises. Investigators made two measurements of temperature rise (°C) on each of 12 such plastic spools. The results are shown below (Nelson, 1986, page 12).

Spool:	1	2	3	4	5	6	7	8	9	10	11	12
First reading:	45.0	45.1	45.4	45.9	45.9	46.0	46.2	46.5	46.5	46.8	47.0	50.6
Second reading:	44.9	44.7	45.8	45.3	45.8	45.2	45.2	45.5	46.0	46.1	45.5	50.0

a. Plot the second reading versus the first reading.

b. Find the least squares line modeling the second reading as a straight-line function of the first reading.

c. Find the least squares line modeling the first reading as a straight-line function of the second reading.

d. Draw the lines you found in parts (b) and (c) on your plot in part (a). Also, draw the standard deviation line. Discuss the meaning of each line.

e. What relationship between the first and second readings would you expect

if the two measurements were consistent? Do the two readings appear to be consistent? Discuss your findings.

EXERCISE 15-11 Researchers sampled 22 naval installations to examine man-hours spent monthly on the clerical task of processing items (Myers, 1986, pages 25–26; from *Procedures and Analyses for Staffing Standards Development: Data/Regression Analysis Handbook,* 1979, Navy Manpower and Material Analysis Center, San Diego, California):

Items processed	Man-hours monthly	Items processed	Man-hours monthly
15	85	527	2,158
25	125	533	2,182
57	203	563	2,302
67	293	563	2,202
197	763	932	3,678
166	639	986	3,894
162	673	1,021	4,034
131	499	1,643	6,622
158	657	1,985	7,890
241	939	1,640	6,610
399	1,546	2,143	8,522

a. Plot monthly man-hours versus items processed.

b. Use the method of least squares to model monthly man-hours as a straight-line function of items processed.

c. Test the null hypothesis that the slope is 0.

d. Test the null hypothesis that the intercept is 0.

e. Use residual plots to check model assumptions.

f. What percentage of the variation in monthly man-hours is explained by the straight-line model? (That is, what is R^2?)

g. Discuss your findings.

EXERCISE 15-12 Investigators studied physical characteristics and ability in 13 American football punters. Each volunteer punted a football ten times. The investigators recorded the average distance for the ten punts, in feet. They also recorded the average hang time (time the ball is in the air before the receiver catches it) for the ten punts, in seconds. In addition, the investigators recorded five measures of strength and/or flexibility for each punter: right leg strength (pounds), left leg strength (pounds), right hamstring muscle flexibility (degrees), left hamstring muscle flexibility (degrees), and overall leg strength (foot-pounds). The results are shown below (Walpole and Myers, 1989, pages 444–445, 450; from the study "The Relationship Between Selected Physical Performance Variables and Football Punting Ability" by the Department of Health, Physical Education, and Recreation at the Virginia Polytechnic Institute and State University, 1983).

Punter	Distance	Hang time	Right leg strength	Left leg strength	Right flexibility	Left flexibility	Overall strength
1	162.50	4.75	170	170	106	106	240.57
2	144.00	4.07	140	130	92	93	195.49
3	147.50	4.04	180	170	93	78	152.99
4	163.50	4.18	160	160	103	93	197.09
5	192.00	4.35	170	150	104	93	266.56
6	171.75	4.16	150	150	101	87	260.56
7	162.00	4.43	170	180	108	106	219.25
8	104.93	3.20	110	110	86	92	132.68
9	105.67	3.02	120	110	90	86	130.24
10	117.59	3.64	130	120	85	80	205.88
11	140.25	3.68	120	140	89	83	153.92
12	150.17	3.60	140	130	92	94	154.64
13	165.17	3.85	160	150	95	95	240.57

a. In this exercise, we will consider only the two variables distance and hang time. Plot the observations of distance and hang time in any ways that seem helpful.

b. Find the least squares line modeling distance as a function of hang time.

c. Find the least squares line modeling hang time as a function of distance.

d. In a scatterplot of distance versus hang time, plot the two lines you found in parts (b) and (c). Also, plot the standard deviation line. Label each line. Discuss the meaning and use of each of these three lines.

EXERCISE 15-13 Refer to the experiment described in Exercise 15-12. For this exercise, consider only the two variables distance and hang time.

a. Plot distance versus hang time.

b. Calculate the linear correlation coefficient for distance and hang time. Test the null hypothesis that the linear correlation coefficient between distance and hang time is 0.

c. Calculate the rank correlation coefficient for distance and hang time. Carry out a nonparametric test that distance and hang time are independent.

d. Compare your results in parts (b) and (c). Discuss your findings.

EXERCISE 15-14 Children with congenital heart defects sometimes need a procedure called heart catheterization. Surgeons pass a 3-mm diameter Teflon tube or catheter into a major vein or artery. They push the tube into the heart to get information on the heart's condition. The surgeons have to guess at the appropriate length of the catheter.

In this study, investigators determined the exact length of the catheter needed in 12 children (Rice, 1988, pages 491–492; from Weindling, 1977). The researchers used a fluoroscope to check when the catheter was in place.

Height, weight, and correct catheter length are shown below for each of the 12 children.

Child	Height (inches)	Weight (pounds)	Catheter length (centimeters)
1	42.8	40.0	37.0
2	63.5	93.5	49.5
3	37.5	35.5	34.5
4	39.5	30.0	36.0
5	45.5	52.0	43.0
6	38.5	17.0	28.0
7	43.0	38.5	37.0
8	22.5	8.5	20.0
9	37.0	33.0	33.5
10	23.5	9.5	30.5
11	33.0	21.0	38.5
12	58.0	79.0	47.0

a. Construct a scatterplot matrix of height, weight, and catheter length.

b. Find the linear correlation coefficient for height and weight, for height and catheter length, and for weight and catheter length.

c. Carry out a simple linear regression analysis, modeling catheter length as a straight-line function of height.

d. Carry out a simple linear regression analysis, modeling catheter length as a straight-line function of weight.

e. Carry out a simple linear regression analysis, modeling weight as a straight-line function of height.

f. Discuss your results. Does it look like surgeons could determine the correct catheter length from the child's height or weight?

EXERCISE 15-15 Refer to the experiment described in Exercise 15-12. For this exercise, consider the variables left leg strength, right leg strength, and distance.

a. Construct a scatterplot matrix of left leg strength, right leg strength, and distance.

b. Find the linear correlation coefficient for left and right leg strength, for left leg strength and distance, and for right leg strength and distance.

c. Carry out a simple linear regression analysis modeling distance as a straight-line function of right leg strength.

d. Carry out a simple linear regression analysis modeling distance as a straight-line function of left leg strength.

e. Carry out a simple linear regression analysis modeling left leg strength as a straight-line function of right leg strength.

f. Discuss your results.

EXERCISE 15-16 In a calibration study in atomic absorption spectroscopy, investigators recorded instrument response in absorbance units at each of five concentrations of copper in solution (Carroll, Sacks, and Spiegelman, 1988):

Run	Amount of copper in solution (micrograms/ milliliter)	Instrument response in absorbance units
1	.0	.045
2	.0	.047
3	.0	.051
4	.0	.054
5	.050	.084
6	.050	.087
7	.100	.115
8	.100	.116
9	.200	.183
10	.200	.191
11	.500	.395
12	.500	.399

a. Plot instrument response versus copper concentration.

b. Use the method of least squares to model instrument response as a straight-line function of copper concentration.

c. Test the null hypothesis that the slope is 0.

d. Test the null hypothesis that the intercept is 0.

e. What percentage of the variation in instrument response is explained by the straight-line model? (That is, what is R^2?)

f. Use residual plots to assess the fit.

g. Discuss your results.

EXERCISE 15-17 Investigators studied the time to failure of samples of electrical insulation for motors in accelerated life testing (Nelson, 1986, pages 20–21). The investigators carried out the accelerated life test at four temperatures, with ten samples of insulation at each temperature.

Temperature (°C)	Hours to failure									
190	7,228	7,228	7,228	8,448	9,167	9,167	9,167	9,167	10,511	10,511
220	1,764	2,436	2,436	2,436	2,436	2,436	3,108	3,108	3,108	3,108
240	1,175	1,175	1,521	1,569	1,617	1,665	1,665	1,713	1,761	1,953
260	600	744	744	744	912	1,128	1,320	1,464	1,608	1,896

a. Plot hours to failure versus temperature.

b. Find the linear correlation coefficient for failure time and temperature.

c. Plot the logarithm of failure time versus the reciprocal of temperature.

d. Find the linear correlation coefficient for the logarithm of failure time and the reciprocal of temperature.

e. Carry out a simple linear regression analysis modeling the logarithm of failure time as a straight-line function of the reciprocal of temperature.

f. Use residual plots to assess the fit of the model in part (e).

g. Discuss your findings.

EXERCISE 15-18 Investigators recorded stopping distance of a car on a road, for several velocities (Rice, 1988, page 505; from Brownlee, 1965, pages 371–372):

Velocity (miles per hour)	Stopping distance (feet)
20.5	15.4
20.5	13.3
30.5	33.9
40.5	73.1
48.8	113.0
57.8	142.6

a. Plot stopping distance versus velocity.

b. Use the method of least squares to model stopping distance as a straight-line function of velocity.

c. What percentage of the variation in stopping distance is explained by the model in part (b)? (That is, what is R^2?)

d. Use residual plots to assess the fit of the model in part (b).

e. Plot the square root of stopping distance versus velocity.

f. Use the method of least squares to model the square root of stopping distance as a straight-line function of velocity.

g. What percentage of the variation in the square root of stopping distance is explained by the model in part (f)? (That is, what is R^2?)

h. Use residual plots to assess the fit of the model in part (f).

i. Discuss your findings.

EXERCISE 15-19 An engineer subjected uniform pieces of stainless steel to different levels of stress, and recorded the time to rupture for each piece. He tested six pieces of steel at each of four stress levels. Stress levels are reported in pounds per square inch (psi). The results are shown below (Schmoyer, 1986; from Garofalo et al., 1961):

Stress level (psi)	Rupture time (hours)					
28.84	1,267	1,637	1,658	1,709	1,785	2,437
31.63	170	257	265	570	594	779
34.68	76	87	96	115	122	132
38.02	22	37	39	41	42	43

a. Plot rupture time versus stress level.

b. Plot the logarithm of rupture time versus stress level.

c. Use the method of least squares to model the logarithm of rupture time as a straight-line function of stress level.

d. Test the null hypothesis that the slope is 0.

e. Test the null hypothesis that the intercept is 0.

f. What percentage of the variation in the logarithm of rupture time is explained by the model in part (c)? (That is, what is R^2?)

g. Use residual plots to check model assumptions.

h. Discuss your findings.

i. Schmoyer (1986) plotted the logarithm of rupture time versus the logarithm of stress level. Construct such a plot. Find $R^2 = r^2$ for these two variables. Compare with your answers to parts (b) and (f).

EXERCISE 15-20 As part of an environmental impact study, investigators looked at the relationship between stream depth and rate of flow (Rice, 1988, page 463; from Ryan, Joiner, and Ryan, 1976). The results are shown below (units not given).

Depth:	.34	.29	.28	.42	.29	.41	.76	.73	.46	.40
Flow rate:	.636	.319	.734	1.327	.487	.924	7.350	5.890	1.979	1.124

a. Plot flow rate versus depth.

b. Use the method of least squares to model flow rate as a straight-line function of depth.

c. Plot residuals versus predicted flow rates for the model in part (b). Does the plot look like random scatter, as it would for an adequate model?

d. Plot the logarithm of flow rate versus the logarithm of depth.

e. Use the method of least squares to model the logarithm of flow rate as a straight-line function of the logarithm of depth.

f. Use residual plots to assess the model in part (e).

g. Discuss your findings.

EXERCISE 15-21 Scientists designed this experiment to study a method for extracting crude oil called the carbon dioxide flooding technique (Mendenhall and Sincich, 1988, page 624; from Wang, 1982). In field use, workers flood carbon dioxide into oil pockets. The carbon dioxide replaces the crude oil, making the oil easier to extract. In this experiment, scientists dipped carbon dioxide flow tubes into oil pockets with known amounts of oil. They tried three carbon dioxide flow pressures and three dipping angles for the flow tubes. The response variable is the percentage of oil recovered. Carbon dioxide flow pressure is recorded in pounds per square inch (psi). The dipping angle is recorded in degrees.

Pressure:	1,000	1,000	1,000	1,500	1,500	1,500	2,000	2,000	2,000
Angle:	0	15	30	0	15	30	0	15	30
Recovery:	60.58	72.72	79.99	66.83	80.78	89.78	69.18	80.31	91.99

a. Construct a scatterplot matrix of these three variables.

b. What do you notice about the relationship between pressure and angle? We could think of this as a two-way factorial experimental design. Each factor (pressure and angle) has three levels. The linear correlation coefficient between pressure and angle is 0. This is a feature of a good experimental design.

c. In the plot(s) containing both pressure and recovery, connect the points with the same values for angle. This creates three profiles, one for each angle. Similarly, in the plot(s) containing both angle and recovery, connect the points with the same values for pressure. This creates three more profiles, one for each pressure. In each of these two plots, are the profiles parallel? Parallel profiles suggest no interaction effect of pressure and angle on recovery. Profiles that are not parallel suggest there is an interaction effect. (For a discussion of interaction effects on a response variable, see Chapter 13.)

d. Use multiple regression methods to model recovery as a function of pressure, angle, and the product of pressure and angle. Use residual plots to assess the fit. Discuss your findings.

e. Use multiple regression methods to model recovery as a function of pressure and angle. Use residual plots to assess the fit. Compare these results with those of part (d).

EXERCISE 15-22 Refer to the experiment described in Exercise 15-12.

a. Use multiple regression to model distance as a function of right leg strength, left leg strength, right flexibility, left flexibility, and overall leg strength.

b. Starting with the model in part (a), go through the steps for a backward regression, stopping when the predictor variable(s) have small p-values.

c. Use residual plots to assess the final model in part (b).

d. Construct a scatterplot matrix of all the variables in the final model in part (b).

e. Discuss your findings.

EXERCISE 15-23 Refer to the experiment described in Exercise 15-12.

a. Use multiple regression to model hang time as a function of right leg strength, left leg strength, right flexibility, left flexibility, and overall leg strength.

b. Starting with the model in part (a), go through the steps for a backward regression, stopping when the predictor variable(s) have small p-values.

c. Use residual plots to assess the final model in part (b).

d. Construct a scatterplot matrix of all the variables in the final model in part (b).

e. Discuss your findings.

EXERCISE 15-24 Researchers conducted this experiment to evaluate the effect of asphalt content on permeability of a type of concrete (Mendenhall and Sincich, 1988, page 495; from Woelfl et al., 1981). They prepared four samples of concrete with each of six levels of asphalt content. They then measured water permeability as the amount of water lost when de-aired water flowed across a sample. Asphalt content is recorded as percentage by total weight of the concrete mix. Permeability is recorded in inches per hour.

Asphalt content	Permeability	Asphalt content	Permeability	Asphalt content	Permeability
3	1,189	5	1,227	7	853
3	840	5	1,180	7	900
3	1,020	5	980	7	733
3	980	5	1,210	7	585
4	1,440	6	707	8	395
4	1,227	6	927	8	270
4	1,022	6	1,067	8	310
4	1,293	6	822	8	208

a. Plot permeability versus asphalt content.

b. Calculate the linear correlation coefficient for permeability and asphalt content. What percentage of the variation in permeability measurements is explained by a straight-line model of permeability as a function of asphalt content?

c. Does it seem reasonable to model permeability as a straight-line function of asphalt content?

d. Use multiple regression to model permeability as a function of asphalt content and the square of asphalt content.

e. What percentage of the variation in permeability measurements is explained by the quadratic model in part (d)? (That is, what is R^2?)

f. Use residual plots to assess the model in part (d).

g. Discuss your findings.