

13

Linear Regression and Correlation

GOALS

When you have completed this chapter, you will be able to:

- 1 Understand and interpret the terms *dependent* and *independent variable*.
- 2 Calculate and interpret the *coefficient of correlation*, the *coefficient of determination*, and the *standard error of estimate*.
- 3 Conduct a test of hypothesis to determine whether the coefficient of correlation in the population is zero.
- 4 Calculate the least squares regression line.
- 5 Construct and interpret confidence and prediction intervals for the dependent variable.



Exercise 61 lists the movies with the largest world box office sales and their world box office budget. Find the correlation between world box office budget and world box office sales. Comment on the association between the two variables. (See Goal 2.)

Introduction



Chapters 2 through 4 dealt with *descriptive statistics*. We organized raw data into a frequency distribution, and computed several measures of location and measures of dispersion to describe the major characteristics of the data. Chapter 5 started the study of *statistical inference*. The main emphasis was on inferring something about a population parameter, such as the population mean, on the basis of a sample. We tested for the reasonableness of a population mean or a population proportion, the difference between two population means, or whether several population means were equal. All of these tests involved just *one* interval- or ratio-level variable, such as the weight of a plastic soft drink bottle, the income of bank presidents, or the number of patients admitted to a particular hospital.

We shift the emphasis in this chapter to the study of two variables. Recall in Chapter 4 we introduced the idea of showing the relationship between two variables with a scatter diagram. We plotted the price of vehicles sold at Whitner Autoplex on the vertical axis and the age of the buyer on the horizontal axis. See the statistical software output on page 119. In that case we observed that, as the age of the buyer increased, the amount spent for the vehicle also increased. In this chapter we carry this idea further. That is, we develop numerical measures to express the relationship between two variables. Is the relationship strong or weak, is it direct or inverse? In addition, we develop an equation to express the relationship between variables. This will allow us to estimate one variable on the basis of another. Here are some examples.

- Is there a relationship between the amount Healthtex spends per month on advertising and its sales in the month?
- Can we base an estimate of the cost to heat a home in January on the number of square feet in the home?
- Is there a relationship between the miles per gallon achieved by large pickup trucks and the size of the engine?
- Is there a relationship between the number of hours that students studied for an exam and the score earned?

Note in each of these cases there are two variables observed for each sampled observation. For the last example, we find, for each student selected for the sample, the hours studied and the score earned.

We begin this chapter by examining the meaning and purpose of **correlation analysis**. We continue our study by developing a mathematical equation that will allow us to estimate the value of one variable based on the value of another. This is called **regression analysis**. We will (1) determine the equation of the line that best fits the data, (2) use the equation to estimate the value of one variable based on another, (3) measure the error in our estimate, and (4) establish confidence and prediction intervals for our estimate.

What Is Correlation Analysis?

Correlation analysis is the study of the relationship between variables. To explain, suppose the sales manager of Copier Sales of America, which has a large sales force throughout the United States and Canada, wants to determine whether there is a relationship between the number of sales calls made in a month and the number of copiers sold that month. The manager selects a random sample of 10 representatives and determines the number of sales calls each representative made



Statistics in Action

The space shuttle Challenger exploded on January 28, 1986. An investigation of the cause examined four contractors: Rockwell International for the shuttle and engines, Lockheed Martin for ground support, Martin Marietta for the external fuel tanks, and Morton Thiokol for the solid fuel booster rockets. After several months, the investigation blamed the explosion on defective O-rings produced by Morton Thiokol. A study of the contractor's stock prices showed an interesting happenstance. On the day of the crash, Morton Thiokol stock was down 11.86% and the stock of the other three lost only 2 to 3%. Can we conclude that financial markets predicted the outcome of the investigation?

Linear Regression and Correlation

459

last month and the number of copiers sold. The sample information is shown in Table 13–1.

TABLE 13–1 Number of Sales Calls and Copiers Sold for 10 Salespeople

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

By reviewing the data we observe that there does seem to be some relationship between the number of sales calls and the number of units sold. That is, the salespeople who made the most sales calls sold the most units. However, the relationship is not “perfect” or exact. For example, Soni Jones made fewer sales calls than Jeff Hall, but she sold more units.

Instead of talking in generalities as we did in Chapter 4 and have so far in this chapter, we will develop some statistical measures to portray more precisely the relationship between the two variables, sales calls and copiers sold. This group of statistical techniques is called **correlation analysis**.

CORRELATION ANALYSIS A group of techniques to measure the association between two variables.

The basic idea of correlation analysis is to report the association between two variables. The usual first step is to plot the data in a **scatter diagram**. An example will show how a scatter diagram is used.

Example

Copier Sales of America sells copiers to businesses of all sizes throughout the United States and Canada. Ms. Marcy Bancer was recently promoted to the position of national sales manager. At the upcoming sales meeting, the sales representatives from all over the country will be in attendance. She would like to impress upon them the importance of making that extra sales call each day. She decides to gather some information on the relationship between the number of sales calls and the number of copiers sold. She selects a random sample of 10 sales representatives and determines the number of sales calls they made last month and the number of copiers they sold. The sample information is reported in Table 13–1. What observations can you make about the relationship between the number of sales calls and the number of copiers sold? Develop a scatter diagram to display the information.

Solution

Based on the information in Table 13–1, Ms. Bancer suspects there is a relationship between the number of sales calls made in a month and the number of copiers sold. Soni Jones sold the most copiers last month, and she was one of three representatives making 30 or more sales calls. On the other hand, Susan Welch and

Carlos Ramirez made only 10 sales calls last month. Ms. Welch, along with two others, had the lowest number of copiers sold among the sampled representatives.

The implication is that the number of copiers sold is related to the number of sales calls made. As the number of sales calls increases, it appears the number of copiers sold also increases. We refer to number of sales calls as the **independent variable** and number of copiers sold as the **dependent variable**.

DEPENDENT VARIABLE The variable that is being predicted or estimated. It is scaled on the Y-axis.

INDEPENDENT VARIABLE The variable that provides the basis for estimation. It is the predictor variable. It is scaled on the X-axis.

It is common practice to scale the dependent variable (copiers sold) on the vertical or Y-axis and the independent variable (number of sales calls) on the horizontal or X-axis. To develop the scatter diagram of the Copier Sales of America sales information, we begin with the first sales representative, Tom Keller. Tom made 20 sales calls last month and sold 30 copiers, so $X = 20$ and $Y = 30$. To plot this point, move along the horizontal axis to $X = 20$, then go vertically to $Y = 30$ and place a dot at the intersection. This process is continued until all the paired data are plotted, as shown in Chart 13–1.

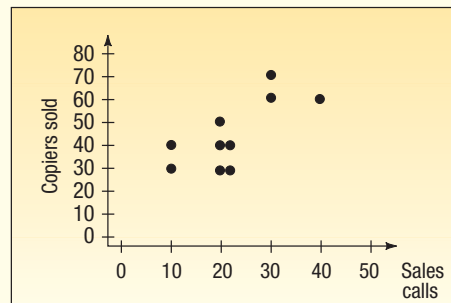


CHART 13–1 Scatter Diagram Showing Sales Calls and Copiers Sold

The scatter diagram shows graphically that the sales representatives who make more calls tend to sell more copiers. It is reasonable for Ms. Bancer, the national sales manager at Copier Sales of America, to tell her salespeople that, the more sales calls they make, the more copiers they can expect to sell. Note that, while there appears to be a positive relationship between the two variables, all the points do not fall on a line. In the following section you will measure the strength and direction of this relationship between two variables by determining the coefficient of correlation.

The Coefficient of Correlation

Interval- or ratio-level
data are required

Characteristics of r

Originated by Karl Pearson about 1900, the **coefficient of correlation** describes the strength of the relationship between two sets of interval-scaled or ratio-scaled variables. Designated r , it is often referred to as *Pearson's r* and as the *Pearson product-moment correlation coefficient*. It can assume any value from -1.00 to $+1.00$ inclusive. A correlation coefficient of -1.00 or $+1.00$ indicates *perfect correlation*. For example, a correlation coefficient for the preceding example computed to be $+1.00$ would indicate that the number of sales calls and the number of copiers sold are perfectly related in a positive linear sense. A computed value of -1.00 reveals that sales calls and the number of copiers sold are

Linear Regression and Correlation

461

perfectly related in an inverse linear sense. How the scatter diagram would appear if the relationship between the two sets of data were linear and perfect is shown in Chart 13–2.

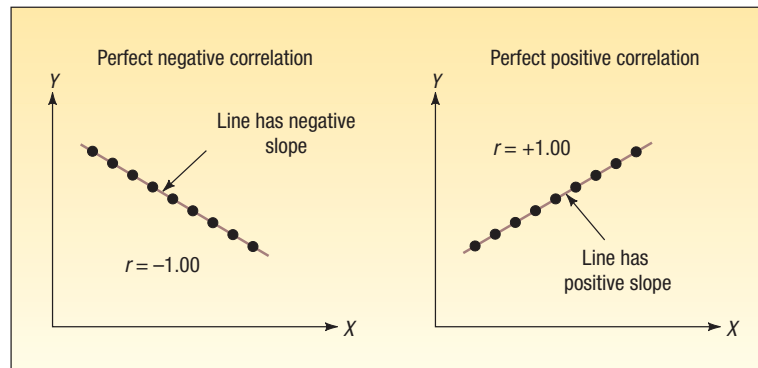


CHART 13–2 Scatter Diagrams Showing Perfect Negative Correlation and Perfect Positive Correlation

If there is absolutely no relationship between the two sets of variables, Pearson's r is zero. A coefficient of correlation r close to 0 (say, .08) shows that the linear relationship is quite weak. The same conclusion is drawn if $r = -.08$. Coefficients of $-.91$ and $+.91$ have equal strength; both indicate very strong correlation between the two variables. Thus, *the strength of the correlation does not depend on the direction (either $-$ or $+$)*.

Scatter diagrams for $r = 0$, a weak r (say, $-.23$), and a strong r (say, $+.87$) are shown in Chart 13–3. Note that, if the correlation is weak, there is considerable scatter about a line drawn through the center of the data. For the scatter diagram representing a strong relationship, there is very little scatter about the line. This indicates, in the example shown on the chart, that hours studied is a good predictor of exam score.

Examples of degrees of
correlation

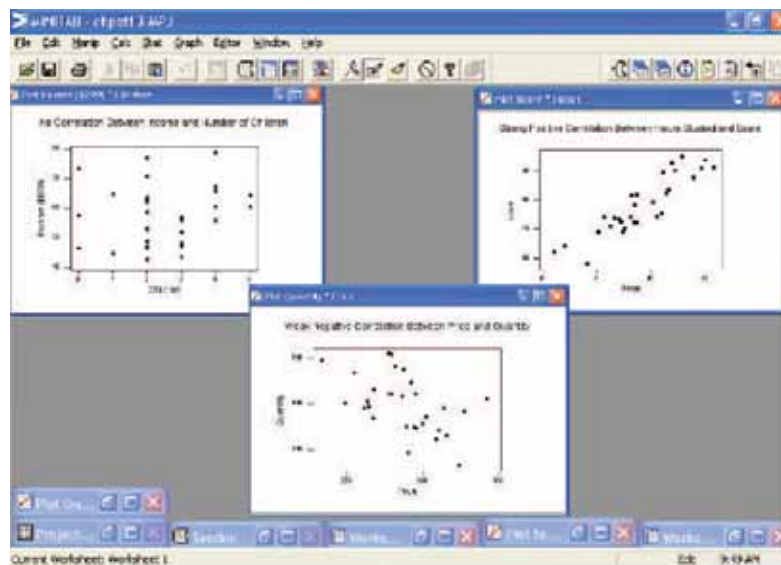
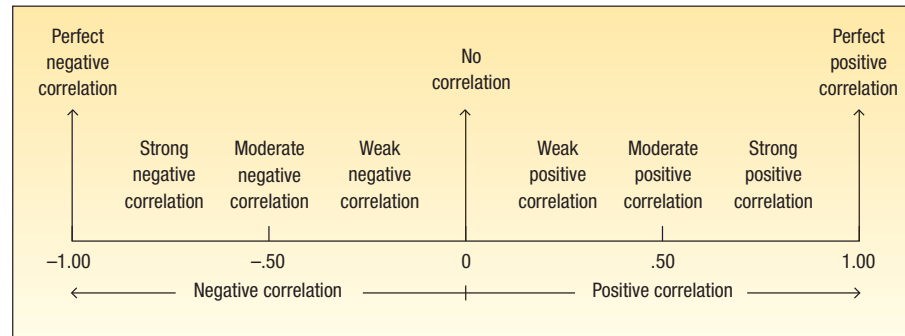


CHART 13–3 Scatter Diagrams Depicting Zero, Weak, and Strong Correlation

Chapter 13

The following drawing summarizes the strength and direction of the coefficient of correlation.



COEFFICIENT OF CORRELATION A measure of the strength of the linear relationship between two variables.

The characteristics of the coefficient of correlation are summarized below.

CHARACTERISTICS OF THE COEFFICIENT OF CORRELATION

1. The sample coefficient of correlation is identified by the lower-case letter r .
2. It shows the direction and strength of the linear (straight line) relationship between two interval- or ratio-scale variables.
3. It ranges from -1 up to and including $+1$.
4. A value near 0 indicates there is little association between the variables.
5. A value near 1 indicates a direct or positive association between the variables.
6. A value near -1 indicates inverse or negative association between the variables.

How is the value of the coefficient of correlation determined? We will use the Copier Sales of America data, which are reported in Table 13–2, as an example. We begin

TABLE 13–2 Sales Calls and Copiers Sold for 10 Salespeople

Sales Representative	Sales Calls, (X)	Copiers Sold, (Y)
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70
Total	220	450

Linear Regression and Correlation

463

with a scatter diagram, similar to Chart 13–2. Draw a vertical line through the data values at the mean of the X -values and a horizontal line at the mean of the Y -values. In Chart 13–4 we've added a vertical line at 22.0 calls ($\bar{X} = \Sigma X/n = 220/10 = 22$) and a horizontal line at 45.0 copiers ($\bar{Y} = \Sigma Y/n = 450/10 = 45.0$). These lines pass through the “center” of the data and divide the scatter diagram into four quadrants. Think of moving the origin from (0, 0) to (22, 45).

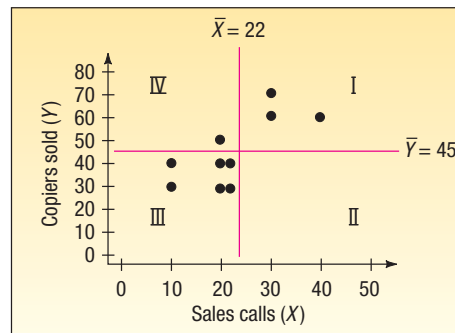


CHART 13–4 Computation of the Coefficient of Correlation

Two variables are positively related when the number of copiers sold is above the mean and the number of sales calls is also above the mean. These points appear in the upper-right quadrant (labeled Quadrant I) of Chart 13–4. Similarly, when the number of copiers sold is less than the mean, so is the number of sales calls. These points fall in the lower-left quadrant of Chart 13–4 (labeled Quadrant III). For example, the last person on the list in Table 13–2, Soni Jones, made 30 sales calls and sold 70 copiers. These values are above their respective means, so this point is located in Quadrant I which is in the upper-right quadrant. She made 8 ($X - \bar{X} = 30 - 22$) more sales calls than the mean and sold 25 ($Y - \bar{Y} = 70 - 45$) more copiers than the mean. Tom Keller, the first name on the list in Table 13–2, made 20 sales calls and sold 30 copiers. Both of these values are less than their respective mean; hence this point is in the lower-left quadrant. Tom made 2 less sales calls and sold 15 less copiers than the respective means. The deviations from the mean number of sales calls and for the mean number of copiers sold are summarized in Table 13–3 for the 10 sales representatives. The sum of the products of the deviations from the respective means is 900. That is, the term $\Sigma(X - \bar{X})(Y - \bar{Y}) = 900$.

In both the upper-right and the lower-left quadrants, the product of $(X - \bar{X})(Y - \bar{Y})$ is positive because both of the factors have the same sign. In our example this

TABLE 13–3 Deviations from the Mean and Their Products

Sales Representative	Calls, X	Sales, Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
Tom Keller	20	30	–2	–15	30
Jeff Hall	40	60	18	15	270
Brian Virost	20	40	–2	–5	10
Greg Fish	30	60	8	15	120
Susan Welch	10	30	–12	–15	180
Carlos Ramirez	10	40	–12	–5	60
Rich Niles	20	40	–2	–5	10
Mike Kiel	20	50	–2	5	–10
Mark Reynolds	20	30	–2	–15	30
Soni Jones	30	70	8	25	200
					<u>900</u>

happens for all sales representatives except Mike Kiel. We can therefore expect the coefficient of correlation to have a positive value.

If the two variables are inversely related, one variable will be above the mean and the other below the mean. Most of the points in this case occur in the upper-left and lower-right quadrants, that is Quadrant II and IV. Now $(X - \bar{X})$ and $(Y - \bar{Y})$ will have opposite signs, so their product is negative. The resulting correlation coefficient is negative.

What happens if there is no linear relationship between the two variables? The points in the scatter diagram will appear in all four quadrants. The negative products of $(X - \bar{X})(Y - \bar{Y})$ offset the positive products, so the sum is near zero. This leads to a correlation coefficient near zero.

The correlation coefficient also needs to be unaffected by the units of the two variables. For example, if we had used hundreds of copiers sold instead of the number sold, the coefficient of correlation would be the same. The coefficient of correlation is independent of the scale used if we divide the term $\Sigma(X - \bar{X})(Y - \bar{Y})$ by the sample standard deviations. It is also made independent of the sample size and bounded by the values $+1.00$ and -1.00 if we divide by $(n - 1)$.

This reasoning leads to the following formula:

CORRELATION COEFFICIENT

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n - 1)s_x s_y} \quad [13-1]$$

To compute the coefficient of correlation, we use the standard deviations of the sample of 10 sales calls and 10 copiers sold. We could use formula (3-12) to calculate the sample standard deviations or we could use a software package. For the specific Excel and MINITAB commands see the **Software Commands** section at the end of Chapter 3. The following is the Excel output. The standard deviation of the number of sales calls is 9.189 and of the number of copiers sold 14.337.



	Calls	Sales		Calls	Sales
1	20	30			
2	40	60	Mean	22.000	Mean 45.000
3	20	40	Standard Error	2.906	Standard Error 4.534
4	30	60	Median	20.000	Median 40.000
5	10	30	Mode	20.000	Mode 30.000
6	10	40	Standard Deviation	9.189	Standard Deviation 14.337
7	20	40	Sample Variance	84.444	Sample Variance 205.556
8	20	60	Kurtosis	0.306	Kurtosis -1.001
9	20	30	Skewness	0.601	Skewness 0.556
10	30	70	Range	30.000	Range 40.000
11			Minimum	10.000	Minimum 30.000
12			Maximum	40.000	Maximum 70.000
13			Sum	220.000	Sum 450.000
14			Count	10.000	Count 10.000

We now insert these values into formula (13-1) to determine the coefficient of correlation:

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n - 1)s_x s_y} = \frac{900}{(10 - 1)(9.189)(14.337)} = 0.759$$

How do we interpret a correlation of 0.759? First, it is positive, so we see there is a direct relationship between the number of sales calls and the number

Linear Regression and Correlation

465

of copiers sold. This confirms our reasoning based on the scatter diagram, Chart 13–4. The value of 0.759 is fairly close to 1.00, so we conclude that the association is strong.

We must be careful with the interpretation. The correlation of 0.759 indicates a strong positive association between the variables. Ms. Bancer would be correct to encourage the sales personnel to make that extra sales call, because the number of sales calls made is *related* to the number of copiers sold. However, does this mean that more sales calls **cause** more sales? No, we have not demonstrated cause and effect here, only that the two variables—sales calls and copiers sold—are related.

The Coefficient of Determination

In the previous example regarding the relationship between the number of sales calls and the units sold, the coefficient of correlation, 0.759, was interpreted as being “strong.” Terms such as *weak*, *moderate*, and *strong*, however, do not have precise meaning. A measure that has a more easily interpreted meaning is the **coefficient of determination**. It is computed by squaring the coefficient of correlation. In the example, the coefficient of determination, r^2 , is 0.576, found by $(0.759)^2$. This is a proportion or a percent; we can say that 57.6 percent of the variation in the number of copiers sold is explained, or accounted for, by the variation in the number of sales calls.

COEFFICIENT OF DETERMINATION The proportion of the total variation in the dependent variable Y that is explained, or accounted for, by the variation in the independent variable X .

Further discussion of the coefficient of determination is found later in the chapter.

Correlation and Cause

If there is a strong relationship (say, .91) between two variables, we are tempted to assume that an increase or decrease in one variable *causes* a change in the other variable. For example, it can be shown that the consumption of Georgia peanuts and the consumption of aspirin have a strong correlation. However, this does not indicate that an increase in the consumption of peanuts *caused* the consumption of aspirin to increase. Likewise, the incomes of professors and the number of inmates in mental institutions have increased proportionately. Further, as the population of donkeys has decreased, there has been an increase in the number of doctoral degrees granted. Relationships such as these are called **spurious correlations**. What we can conclude when we find two variables with a strong correlation is that there is a relationship or association between the two variables, not that a change in one causes a change in the other.

Self-Review 13–1



Haverty's Furniture is a family business that has been selling to retail customers in the Chicago area for many years. The company advertises extensively on radio, TV, and the Internet, emphasizing low prices and easy credit terms. The owner would like to review the relationship between sales and the amount spent on advertising. Below is information on sales and advertising expense for the last four months.

Month	Advertising Expense (\$ million)	Sales Revenue (\$ million)
July	2	7
August	1	3
September	3	8
October	4	10

- (a) The owner wants to forecast sales on the basis of advertising expense. Which variable is the dependent variable? Which variable is the independent variable?

- (b) Draw a scatter diagram.
- (c) Determine the coefficient of correlation.
- (d) Interpret the strength of the correlation coefficient.
- (e) Determine the coefficient of determination. Interpret.

Exercises

1. The following sample observations were randomly selected.

X:	4	5	3	6	10
Y:	4	6	5	7	7

Determine the coefficient of correlation and the coefficient of determination. Interpret.

2. The following sample observations were randomly selected.

X:	5	3	6	3	4	4	6	8
Y:	13	15	7	12	13	11	9	5

Determine the coefficient of correlation and the coefficient of determination. Interpret the association between X and Y.

3. Bi-lo Appliance Super-Store has outlets in several large metropolitan areas in New England. The general sales manager plans to air a commercial for a digital camera on selected local TV stations prior to a sale starting on Saturday and ending Sunday. She plans to get the information for Saturday–Sunday digital camera sales at the various outlets and pair them with the number of times the advertisement was shown on the local TV stations. The purpose is to find whether there is any relationship between the number of times the advertisement was aired and digital camera sales. The pairings are:

Location of TV Station	Number of Airings	Saturday–Sunday Sales (\$ thousands)
Providence	4	15
Springfield	2	8
New Haven	5	21
Boston	6	24
Hartford	3	17

- a. What is the dependent variable?
- b. Draw a scatter diagram.
- c. Determine the coefficient of correlation.
- d. Determine the coefficient of determination.
- e. Interpret these statistical measures.
4. The production department of Celltronics International wants to explore the relationship between the number of employees who assemble a subassembly and the number produced. As an experiment, two employees were assigned to assemble the subassemblies. They produced 15 during a one-hour period. Then four employees assembled them. They produced 25 during a one-hour period. The complete set of paired observations follows.

Number of Assemblers	One-Hour Production (units)
2	15
4	25
1	10
5	40
3	30

Linear Regression and Correlation

467

The dependent variable is production; that is, it is assumed that the level of production depends upon the number of employees.

- Draw a scatter diagram.
 - Based on the scatter diagram, does there appear to be any relationship between the number of assemblers and production? Explain.
 - Compute the coefficient of correlation.
 - Evaluate the strength of the relationship by computing the coefficient of determination.
5. The city council of Pine Bluffs is considering increasing the number of police in an effort to reduce crime. Before making a final decision, the council asks the chief of police to survey other cities of similar size to determine the relationship between the number of police and the number of crimes reported. The chief gathered the following sample information.

City	Police	Number of Crimes	City	Police	Number of Crimes
Oxford	15	17	Holgate	17	7
Starksville	17	13	Carey	12	21
Danville	25	5	Whistler	11	19
Athens	27	7	Woodville	22	6

- If we want to estimate crimes on the basis of the number of police, which variable is the dependent variable and which is the independent variable?
 - Draw a scatter diagram.
 - Determine the coefficient of correlation.
 - Determine the coefficient of determination.
 - Interpret these statistical measures. Does it surprise you that the relationship is inverse?
6. The owner of Maumee Ford-Mercury-Volvo wants to study the relationship between the age of a car and its selling price. Listed below is a random sample of 12 used cars sold at the dealership during the last year.

Car	Age (years)	Selling Price (\$000)	Car	Age (years)	Selling Price (\$000)
1	9	8.1	7	8	7.6
2	7	6.0	8	11	8.0
3	11	3.6	9	10	8.0
4	12	4.0	10	12	6.0
5	8	5.0	11	6	8.6
6	7	10.0	12	6	8.0

- If we want to estimate selling price on the basis of the age of the car, which variable is the dependent variable and which is the independent variable?
- Draw a scatter diagram.
- Determine the coefficient of correlation.
- Determine the coefficient of determination.
- Interpret these statistical measures. Does it surprise you that the relationship is inverse?

Testing the Significance of the Correlation Coefficient

Recall that the sales manager of Copier Sales of America found the correlation between the number of sales calls and the number of copiers sold was 0.759. This indicated a strong association between the two variables. However, only 10 salespeople were sampled. Could it be that the correlation in the population is actually 0? This would mean the correlation of 0.759 was due to chance. The population in this example is all the salespeople employed by the firm.

Resolving this dilemma requires a test to answer the obvious question: Could there be zero correlation in the population from which the sample was selected? To put it another way, did the computed r come from a population of paired

Could the correlation in the population be zero?

Chapter 13

observations with zero correlation? To continue our convention of allowing Greek letters to represent a population parameter, we will let ρ represent the correlation in the population. It is pronounced “rho.”

We will continue with the illustration involving sales calls and copiers sold. We employ the same hypothesis testing steps described in Chapter 10. The null hypothesis and the alternate hypothesis are:

$$\begin{aligned} H_0: \rho &= 0 && \text{(The correlation in the population is zero.)} \\ H_1: \rho &\neq 0 && \text{(The correlation in the population is different from zero.)} \end{aligned}$$

From the way H_1 is stated, we know that the test is two-tailed.

The formula for t is:

 **t TEST FOR THE
COEFFICIENT OF
CORRELATION**

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

with $n - 2$ degrees of freedom

[13–2]

Using the .05 level of significance, the decision rule in this instance states that if the computed t falls in the area between plus 2.306 and minus 2.306, the null hypothesis is not rejected. To locate the critical value of 2.306, refer to Appendix B.2 for $df = n - 2 = 10 - 2 = 8$. See Chart 13–5.

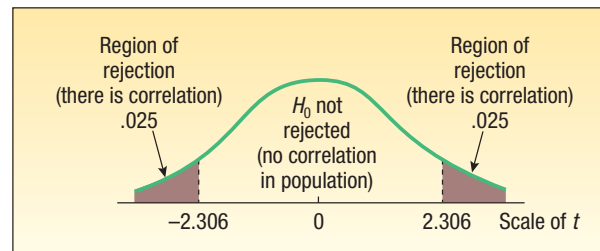


CHART 13–5 Decision Rule for Test of Hypothesis at .05 Significance Level and 8 df

Applying formula (13–2) to the example regarding the number of sales calls and units sold:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.759\sqrt{10-2}}{\sqrt{1-.759^2}} = 3.297$$

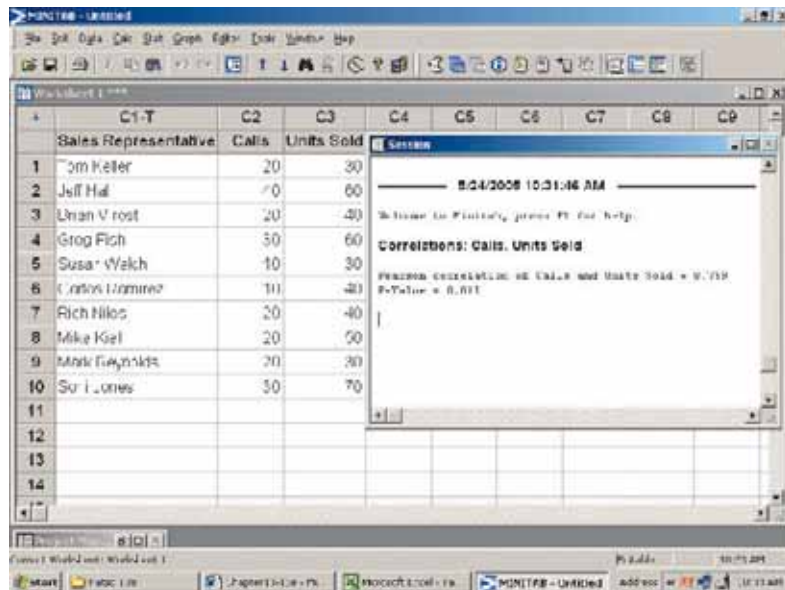
The computed t is in the rejection region. Thus, H_0 is rejected at the .05 significance level. This means the correlation in the population is not zero. From a practical standpoint, it indicates to the sales manager that there is correlation with respect to the number of sales calls made and the number of copiers sold in the population of salespeople.

We can also interpret the test of hypothesis in terms of p -values. A p -value is the likelihood of finding a value of the test statistic more extreme than the one computed, when H_0 is true. To determine the p -value, go to the t distribution in Appendix B.2 and find the row for 8 degrees of freedom. The value of the test statistic is 3.297, so in the row for 8 degrees of freedom and a two-tailed test, find the value closest to 3.297. For a two-tailed test at the .02 significance level, the critical value is 2.896, and the critical value at the .01 significance level is 3.355. Because 3.297 is between 2.896 and 3.355 we conclude that the p -value is between .01 and .02.

Both MINITAB and Excel will report the correlation between two variables. In addition to the correlation, MINITAB reports the p -value for the test of hypothesis that the correlation in the population between the two variables is 0. The MINITAB output showing the results is below. They are the same as those calculated earlier.

Linear Regression and Correlation

469



Self-Review 13–2



A sample of 25 mayoral campaigns in medium-sized cities with populations between 50,000 and 250,000 showed that the correlation between the percent of the vote received and the amount spent on the campaign by the candidate was .43. At the .05 significance level, is there a positive association between the variables?

Exercises

7. The following hypotheses are given.

$$H_0: \rho \leq 0$$

$$H_1: \rho > 0$$

A random sample of 12 paired observations indicated a correlation of .32. Can we conclude that the correlation in the population is greater than zero? Use the .05 significance level.

8. The following hypotheses are given.

$$H_0: \rho \geq 0$$

$$H_1: \rho < 0$$

A random sample of 15 paired observations have a correlation of $-.46$. Can we conclude that the correlation in the population is less than zero? Use the .05 significance level.

9. Pennsylvania Refining Company is studying the relationship between the pump price of gasoline and the number of gallons sold. For a sample of 20 stations last Tuesday, the correlation was .78. At the .01 significance level, is the correlation in the population greater than zero?
10. A study of 20 worldwide financial institutions showed the correlation between their assets and pretax profit to be .86. At the .05 significance level, can we conclude that there is positive correlation in the population?
11. The Airline Passenger Association studied the relationship between the number of passengers on a particular flight and the cost of the flight. It seems logical that more passengers on the flight will result in more weight and more luggage, which in turn will result in higher fuel costs. For a sample of 15 flights, the correlation between the number of passengers and total fuel cost was .667. Is it reasonable to conclude that there is positive association in the population between the two variables? Use the .01 significance level.

12. The Student Government Association at Middle Carolina University wanted to demonstrate the relationship between the number of beers a student drinks and their blood alcohol content (BAC). A random sample of 18 students participated in a study in which each participating student was randomly assigned a number of 12-ounce cans of beer to drink. Thirty minutes after consuming their assigned number of beers a member of the local sheriff's office measured their blood alcohol content. The sample information is reported below.

Student	Beers	BAC	Student	Beers	BAC
1	6	0.10	10	3	0.07
2	7	0.09	11	3	0.05
3	7	0.09	12	7	0.08
4	4	0.10	13	1	0.04
5	5	0.10	14	4	0.07
6	3	0.07	15	2	0.06
7	3	0.10	16	7	0.12
8	6	0.12	17	2	0.05
9	6	0.09	18	1	0.02

Use a statistical software package to answer the following questions.

- Develop a scatter diagram for the number of beers consumed and BAC. Comment on the relationship. Does it appear to be strong or weak? Does it appear to be direct or inverse?
- Determine the coefficient of correlation.
- Determine the coefficient of determination.
- At the .01 significance level is it reasonable to conclude that there is a positive relationship in the population between the number of beers consumed and the BAC? What is the p -value?

Regression Analysis



In the previous section we developed measures to express the strength and the direction of the linear relationship between two variables. In this section we wish to develop an equation to express the *linear* (straight line) relationship between two variables. In addition we want to be able to estimate the value of the dependent variable Y based on a selected value of the independent variable X . The technique used to develop the equation and provide the estimates is called **regression analysis**.

In Table 13–1 we reported the number of sales calls and the number of units sold for a sample of 10 sales representatives employed by Copier Sales of America. Chart 13–1 portrayed this information in a scatter diagram. Now we want to develop a linear equation that expresses the relationship between the number of sales calls and the number of units sold. The equation for the line used to estimate Y on the basis of X is referred to as the **regression equation**.

REGRESSION EQUATION An equation that expresses the linear relationship between two variables.

Least Squares Principle

The scatter diagram in Chart 13–1 is reproduced in Chart 13–6, with a line drawn with a ruler through the dots to illustrate that a straight line would probably fit the data. However, the line drawn using a straight edge has one disadvantage: Its position is based in part on the judgment of the person drawing the line. The hand-

Linear Regression and Correlation

471

drawn lines in Chart 13–7 represent the judgments of four people. All the lines except line A seem to be reasonable. However, each would result in a different estimate of units sold for a particular number of sales calls.

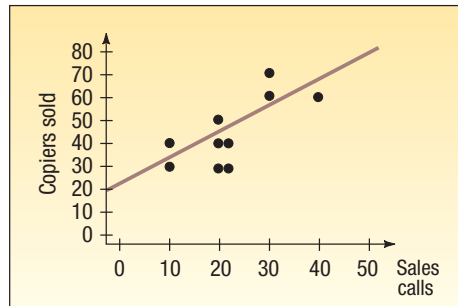


CHART 13–6 Sales Calls and Copiers Sold
for 10 Sales Representatives

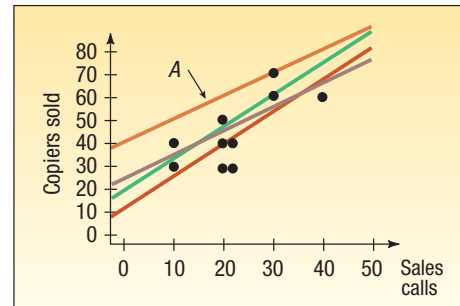


CHART 13–7 Four Lines Superimposed on
the Scatter Diagram

Least squares line gives
“best” fit; subjective
method is unreliable

Judgment is eliminated by determining the regression line using a mathematical method called the **least squares principle**. This method gives what is commonly referred to as the “best-fitting” line.

LEAST SQUARES PRINCIPLE Determining a regression equation by minimizing the sum of the squares of the vertical distances between the actual Y values and the predicted values of Y .

To illustrate this concept, the same data are plotted in the three charts that follow. The regression line in Chart 13–8 was determined using the least squares method. It is the best-fitting line because the sum of the squares of the vertical deviations about it is at a minimum. The first plot ($X = 3$, $Y = 8$) deviates by 2 from the line, found by $10 - 8$. The deviation squared is 4. The squared deviation for the plot $X = 4$, $Y = 18$ is 16. The squared deviation for the plot $X = 5$, $Y = 16$ is 4. The sum of the squared deviations is 24, found by $4 + 16 + 4$.

Assume that the lines in Charts 13–9 and 13–10 were drawn with a straight edge. The sum of the squared vertical deviations in Chart 13–9 is 44. For Chart 13–10

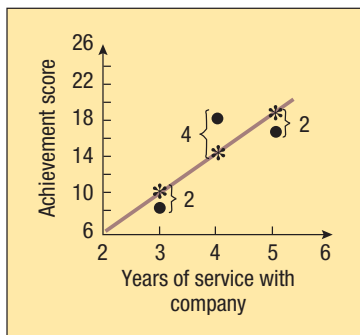


CHART 13–8 The Least Squares
Line

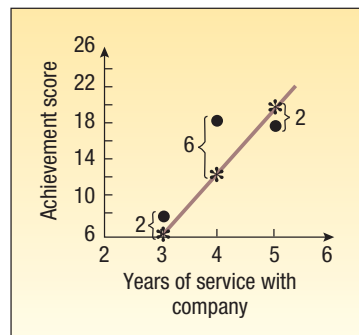


CHART 13–9 Line Drawn with a
Straight Edge

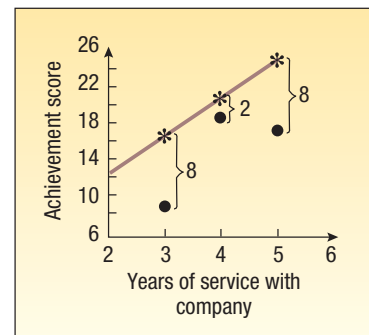


CHART 13–10 Different Line Drawn
with a Straight Edge

Chapter 13

it is 132. Both sums are greater than the sum for the line in Chart 13–8, found by using the least squares method.

The equation of a straight line has the form

$$\text{GENERAL FORM OF LINEAR REGRESSION EQUATION} \quad \hat{Y} = a + bX \quad [13-3]$$

where

\hat{Y} , read Y hat, is the estimated value of the Y variable for a selected X value.
 a is the Y -intercept. It is the estimated value of Y when $X = 0$. Another way to put it is: a is the estimated value of Y where the regression line crosses the Y -axis when X is zero.

b is the slope of the line, or the average change in \hat{Y} for each change of one unit (either increase or decrease) in the independent variable X .

X is any value of the independent variable that is selected.

The formulas for a and b are:

$$\text{SLOPE OF THE REGRESSION LINE} \quad b = r \frac{s_y}{s_x} \quad [13-4]$$

where

r is the correlation coefficient.

s_y is the standard deviation of Y (the dependent variable).

s_x is the standard deviation of X (the independent variable).

$$\text{Y-INTERCEPT} \quad a = \bar{Y} - b\bar{X} \quad [13-5]$$

where

\bar{Y} is the mean of Y (the dependent variable).

\bar{X} is the mean of X (the independent variable).

Example

Recall the example involving Copier Sales of America. The sales manager gathered information on the number of sales calls made and the number of copiers sold for a random sample of 10 sales representatives. As a part of her presentation at the upcoming sales meeting, Ms. Bancer, the sales manager, would like to offer specific information about the relationship between the number of sales calls and the number of copiers sold. Use the least squares method to determine a linear equation to express the relationship between the two variables. What is the expected number of copiers sold by a representative who made 20 calls?

Solution

The first step in determining the regression equation is to find the slope of the least squares regression line. That is, we need the value of b . On page 464, we determined the correlation coefficient r (.759). In the Excel output on page 464, we determined the standard deviation of the independent variable X (9.189) and the standard deviation of the dependent variable Y (14.337). The values are inserted in formula 13–4.

$$b = r \left(\frac{s_y}{s_x} \right) = .759 \left(\frac{14.337}{9.189} \right) = 1.1842$$

Next we need to find the value of a . To do this we use the value for b that we just calculated as well as the means for the number of sales calls and the number of

Linear Regression and Correlation

473

copiers sold. These means are also available in the Excel printout on page 464. From formula (13–5):

$$a = \bar{Y} - b\bar{X} = 45 - 1.1842(22) = 18.9476$$

Thus, the regression equation is $\hat{Y} = 18.9476 + 1.1842X$. So if a salesperson makes 20 calls, he or she can expect to sell 42.6316 copiers, found by $\hat{Y} = 18.9476 + 1.1842X = 18.9476 + 1.1842(20)$. The b value of 1.1842 means that for each additional sales call made the sales representative can expect to increase the number of copiers sold by about 1.2. To put it another way, five additional sales calls in a month will result in about six more copiers being sold, found by $1.1842(5) = 5.921$.

The a value of 18.9476 is the point where the equation crosses the Y -axis. A literal translation is that if no sales calls are made, that is, $X = 0$, 18.9476 copiers will be sold. Note that $X = 0$ is outside the range of values included in the sample and, therefore, should not be used to estimate the number of copiers sold. The sales calls ranged from 10 to 40, so estimates should be made within that range.

**Statistics in Action**

In finance, investors are interested in the trade-off between returns and risk. One technique to quantify risk is a regression analysis of a company's stock price (dependent variable) and an average measure of the stock market (independent variable). Often the Standard and Poor's (S&P) 500 Index is used to estimate the market. The regression coefficient, called beta in finance, shows the change in a company's stock price for a one-unit change in the S&P Index. For example, if a stock has a beta of 1.5, then when the S&P index

(continued)

Drawing the Regression Line

The least squares equation, $\hat{Y} = 18.9476 + 1.1842X$, can be drawn on the scatter diagram. The first sales representative in the sample is Tom Keller. He made 20 calls. His estimated number of copiers sold is $\hat{Y} = 18.9476 + 1.1842(20) = 42.6316$. The plot $X = 20$ and $Y = 42.6316$ is located by moving to 20 on the X -axis and then going vertically to 42.6316. The other points on the regression equation can be determined by substituting the particular value of X into the regression equation.

Sales Representative	Sales Calls (X)	Estimated Sales (Ŷ)	Sales Representative	Sales Calls (X)	Estimated Sales (Ŷ)
Tom Keller	20	42.6316	Carlos Ramirez	10	30.7896
Jeff Hall	40	66.3156	Rich Niles	20	42.6316
Brian Virost	20	42.6316	Mike Kiel	20	42.6316
Greg Fish	30	54.4736	Mark Reynolds	20	42.6316
Susan Welch	10	30.7896	Soni Jones	30	54.4736

All the other points are connected to give the line. See Chart 13–11.

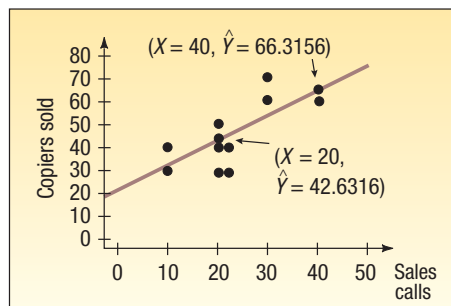


CHART 13–11 The Line of Regression Drawn on the Scatter Diagram

increases by 1%, the stock price will increase by 1.5%. The opposite is also true. If the S&P decreases by 1%, the stock price will decrease by 1.5%. If the beta is 1.0, then a 1% change in the index should show a 1% change in a stock price. If the beta is less than 1.0, then a 1% change in the index shows less than a 1% change in the stock price.



The least squares regression line has some interesting and unique features. First, it will always pass through the point (\bar{X}, \bar{Y}) . To show this is true, we can use the mean number of sales calls to predict the number of copiers sold. In this example the mean number of sales calls is 22.0, found by $\bar{X} = 220/10$. The mean number of copiers sold is 45.0, found by $\bar{Y} = 450/10 = 45$. If we let $X = 22$ and then use the regression equation to find the estimated value for \hat{Y} , the result is:

$$\hat{Y} = 18.9476 + 1.1842 \times 22 = 45$$

The estimated number of copiers sold is exactly equal to the mean number of copiers sold. This simple example shows the regression line will pass through the point represented by the two means. In this case the regression equation will pass through the point $X = 22$ and $Y = 45$.

Second, as we discussed earlier in this section, there is no other line through the data where the sum of the squared deviations is smaller. To put it another way, the term $\Sigma(Y - \hat{Y})^2$ is smaller for the least squares regression equation than for any other equation. We use the Excel system to demonstrate this condition.

Sales Representative	Sales Calls (X)	Copiers Sold (Y)	Estimated Sales (\hat{Y})	Residuals ($Y - \hat{Y}$)	Squared Residuals ($(Y - \hat{Y})^2$)
Tom Miller	20	30	42.6316	-12.6316	159.5573
Jerry Bell	40	60	66.2136	-6.2136	38.6088
Wendy Vogel	20	40	42.6316	-2.6316	6.9275
Gary Fox	30	50	54.4736	-5.5264	30.5409
Paul Woods	10	20	30.7896	-10.7896	116.4351
Cathy Roberts	10	40	30.7896	9.2104	84.8318
Rich Miller	20	30	42.6316	-12.6316	159.5573
Mike Hall	20	60	42.6316	17.3684	301.6318
Mark Reynolds	20	50	42.6316	7.3684	54.2933
Soni Jones	30	70	54.4736	15.5264	241.0691
Totals	220	450	450	0	784.2105

In Columns A, B, and C in the Excel spreadsheet above we duplicated the sample information on sales and copiers sold from Table 13-1. In column D we provide the estimated sales values, the \hat{Y} values, as calculated above.

In column E we calculate the **residuals**, or the error values. This is the difference between the actual values and the predicted values. That is, column E is $(Y - \hat{Y})$. For Soni Jones,

$$\hat{Y} = 18.9476 + 1.1842 \times 30 = 54.4736$$

Her actual value is 70. So the residual, or error of estimate, is

$$(Y - \hat{Y}) = (70 - 54.4736) = 15.5264$$

This value reflects the amount the predicted value of sales is “off” from the actual sales value.

Next in Column F we square the residuals for each of the sales representatives and total the result. The total is 784.2105.

$$\Sigma(Y - \hat{Y})^2 = 159.5573 + 39.8868 + \cdots + 241.0691 = 784.2105$$

This is the sum of the squared differences or the least squares value. There is no other line through these 10 data points where the sum of the squared differences is smaller.

Linear Regression and Correlation

475

We can demonstrate the least squares criterion by choosing two arbitrary equations that are close to the least squares equation and determining the sum of the squared differences for these equations. In column G we use the equation $Y^* = 19 + 1.2X$ to find the predicted value. Notice this equation is very similar to the least squares equation. In Column H we determine the residuals and square these residuals. For the first sales representative, Tom Keller,

$$Y^* = 19 + 1.2(20) = 43$$

$$(Y - Y^*)^2 = (43 - 30)^2 = 169$$

This procedure is continued for the other nine sales representatives and the squared residuals totaled. The result is 786. This is a larger value (786 versus 784.2105) than the residuals for the least squares line.

In columns I and J on the output we repeat the above process for yet another equation $Y^{**} = 20 + X$. Again, this equation is similar to the least squares equation. The details for Tom Keller are:

$$Y^{**} = 20 + X = 20 + 20 = 40$$

$$(Y - Y^{**})^2 = (30 - 40)^2 = 100$$

This procedure is continued for the other nine sales representatives and the residuals totaled. The result is 900, which is also larger than the least squares values.

What have we shown with the example? The sum of the squared residuals ($\sum(Y - \hat{Y})^2$) for the least squares equation is smaller than for other selected lines. The bottom line is you will not be able to find a line passing through these data points where the sum of the squared residuals is smaller.

Self-Review 13–3



Refer to Self-Review 13–1, where the owner of Haverty's Furniture Company was studying the relationship between sales and the amount spent on advertising. The sales information for the last four months is repeated below.

Month	Advertising Expense (\$ million)	Sales Revenue (\$ million)
July	2	7
August	1	3
September	3	8
October	4	10

- Determine the regression equation.
- Interpret the values of a and b .
- Estimate sales when \$3 million is spent on advertising.

Exercises

13. The following sample observations were randomly selected.

X:	4	5	3	6	10
Y:	4	6	5	7	7

- Determine the regression equation.
 - Determine the value of \hat{Y} when X is 7.
14. The following sample observations were randomly selected.

X:	5	3	6	3	4	4	6	8
Y:	13	15	7	12	13	11	9	5

Chapter 13

- a. Determine the regression equation.
b. Determine the value of \hat{Y} when X is 7.
15. Bradford Electric Illuminating Company is studying the relationship between kilowatt-hours (thousands) used and the number of rooms in a private single-family residence. A random sample of 10 homes yielded the following.

Number of Rooms	Kilowatt-Hours (thousands)	Number of Rooms	Kilowatt-Hours (thousands)
12	9	8	6
9	7	10	8
14	10	10	10
6	5	5	4
10	8	7	7

- a. Determine the regression equation.
b. Determine the number of kilowatt-hours, in thousands, for a six-room house.
16. Mr. James McWhinney, president of Daniel-James Financial Services, believes there is a relationship between the number of client contacts and the dollar amount of sales. To document this assertion, Mr. McWhinney gathered the following sample information. The X column indicates the number of client contacts last month, and the Y column shows the value of sales (\$ thousands) last month for each client sampled.

Number of Contacts, X	Sales (\$ thousands), Y	Number of Contacts, X	Sales (\$ thousands), Y
14	24	23	30
12	14	48	90
20	28	50	85
16	30	55	120
46	80	50	110

- a. Determine the regression equation.
b. Determine the estimated sales if 40 contacts are made.
17. A recent article in *BusinessWeek* listed the “Best Small Companies.” We are interested in the current results of the companies’ sales and earnings. A random sample of 12 companies was selected and the sales and earnings, in millions of dollars, are reported below.

Company	Sales (\$ millions)	Earnings (\$ millions)	Company	Sales (\$ millions)	Earnings (\$ millions)
Papa John's International	\$89.2	\$4.9	Checkmate Electronics	\$17.5	\$ 2.6
Applied Innovation	18.6	4.4	Royal Grip	11.9	1.7
Integracare	18.2	1.3	M-Wave	19.6	3.5
Wall Data	71.7	8.0	Serving-N-Slide	51.2	8.2
Davidson & Associates	58.6	6.6	Daig	28.6	6.0
Chico's FAS	46.8	4.1	Cobra Golf	69.2	12.8

Let sales be the independent variable and earnings be the dependent variable.

- a. Draw a scatter diagram.
b. Compute the coefficient of correlation.
c. Compute the coefficient of determination.
d. Interpret your findings in parts (b) and (c).
e. Determine the regression equation.
f. For a small company with \$50.0 million in sales, estimate the earnings.
18. We are studying mutual bond funds for the purpose of investing in several funds. For this particular study, we want to focus on the assets of a fund and its five-year performance. The question is: Can the five-year rate of return be estimated based on the assets of the

Linear Regression and Correlation

477

fund? Nine mutual funds were selected at random, and their assets and rates of return are shown below.

Fund	Assets (\$ millions)	Return (%)	Fund	Assets (\$ millions)	Return (%)
AARP High Quality Bond	\$622.2	10.8	MFS Bond A	\$494.5	11.6
Babson Bond L	160.4	11.3	Nichols Income	158.3	9.5
Compass Capital Fixed Income	275.7	11.4	T. Rowe Price Short-term	681.0	8.2
Galaxy Bond Retail	433.2	9.1	Thompson Income B	241.3	6.8
Keystone Custodian B-1	437.9	9.2			

- Draw a scatter diagram.
 - Compute the coefficient of correlation.
 - Compute the coefficient of determination.
 - Write a brief report of your findings for parts (b) and (c).
 - Determine the regression equation. Use assets as the independent variable.
 - For a fund with \$400.0 million in sales, determine the five-year rate of return (in percent).
19. Refer to Exercise 5.
- Determine the regression equation.
 - Estimate the number of crimes for a city with 20 police.
 - Interpret the regression equation.
20. Refer to Exercise 6.
- Determine the regression equation.
 - Estimate the selling price of a 10-year-old car.
 - Interpret the regression equation.

The Standard Error of Estimate

Note in the preceding scatter diagram (Chart 13–11) that all of the points do not lie exactly on the regression line. If they all were on the line, there would be no error in estimating the number of units sold. To put it another way, if all the points were on the regression line, units sold could be predicted with 100 percent accuracy. Thus, there would be no error in predicting the Y variable based on an X variable. This is true in the following hypothetical case (see Chart 13–12). Theoretically, if $X = 4$, then an exact Y of 100 could be predicted with 100 percent confidence. Or if $X = 12$, then $Y = 300$. Because there is no difference between the observed values and the predicted values, there is no error in this estimate.

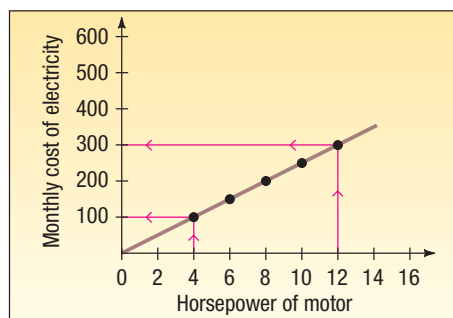


CHART 13–12 Example of Perfect Prediction: Horsepower and Cost of Electricity

Perfect prediction
unrealistic in business

Perfect prediction in economics and business is practically impossible. For example, the revenue for the year from gasoline sales (Y) based on the number of

automobile registrations (X) as of a certain date could no doubt be approximated fairly closely, but the prediction would not be exact to the nearest dollar, or probably even to the nearest thousand dollars. Even predictions of tensile strength of steel wires based on the outside diameters of the wires are not always exact due to slight differences in the composition of the steel.

What is needed, then, is a measure that describes how precise the prediction of Y is based on X or, conversely, how inaccurate the estimate might be. This measure is called the **standard error of estimate**. The standard error of estimate, symbolized by $s_{y \cdot x}$, is the same concept as the standard deviation discussed in Chapter 3. The standard deviation measures the dispersion around the mean. The standard error of estimate measures the dispersion about the regression line.

STANDARD ERROR OF ESTIMATE A measure of the dispersion, or scatter, of the observed values around the line of regression.

The standard error of estimate is found using the following formula, (13–6). Note the following important features:

1. It is similar to the standard deviation in that it is based on squared deviations. The numerator of the standard deviation as we calculated in formula (3–11) on page 79 is based on squared deviations from the mean. The numerator of the standard error is based on squared deviations from the regression line.
2. The sum of the squared deviations is the least squares value used to find the best fitting regression line. Recall in the previous section we described how to find the least squares value (see Column F of the Excel spreadsheet on page 474). We compared the least squares value to values generated from other lines plotted through the data.
3. The denominator of the equation is $n - 2$. As usual, n is the number of observations. We lose two degrees of freedom because we are estimating two parameters. So the values of b , the slope of the line, and a , the Y -intercept, are sample values we use to estimate their corresponding population values. We are sampling from a population and are estimating the slope of the line and the intercept with the Y -axis. Hence the denominator is $n - 2$.

STANDARD ERROR OF ESTIMATE

$$s_{y \cdot x} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n - 2}} \quad [13-6]$$

If $s_{y \cdot x}$ is small, this means that the data are relatively close to the regression line and the regression equation can be used to predict Y with little error. If $s_{y \cdot x}$ is large, this means that the data are widely scattered around the regression line and the regression equation will not provide a precise estimate Y .

Example

Recall the example involving Copier Sales of America. The sales manager determined the least squares regression equation to be $\hat{Y} = 18.9476 + 1.1842X$, where \hat{Y} refers to the predicted number of copiers sold and X the number of sales calls made. Determine the standard error of estimate as a measure of how well the values fit the regression line.

Solution

To find the standard error, we begin by finding the difference between the value, Y , and the value estimated from the regression equation, \hat{Y} . Next we square this difference, that is, $(Y - \hat{Y})^2$. We do this for each of the n observations and sum the

Linear Regression and Correlation

479

results. That is, we compute $\Sigma(Y - \hat{Y})^2$, which is the numerator of formula (13–6). Finally, we divide by the number of observations minus 2. The details of the calculations are summarized in Table 13–4.

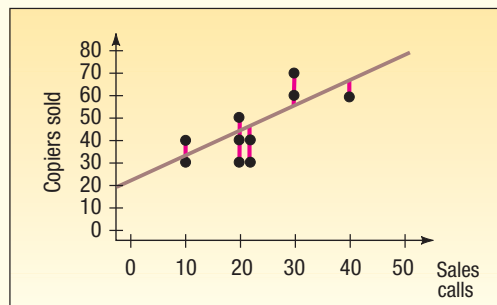
TABLE 13–4 Computations Needed for the Standard Error of Estimate

Sales Representative	Actual Sales, (Y)	Estimated Sales, (\hat{Y})	Deviation, (Y – \hat{Y})	Deviation Squared, (Y – \hat{Y}) ²
Tom Keller	30	42.6316	–12.6316	159.557
Jeff Hall	60	66.3156	–6.3156	39.887
Brian Virost	40	42.6316	–2.6316	6.925
Greg Fish	60	54.4736	5.5264	30.541
Susan Welch	30	30.7896	–0.7896	0.623
Carlos Ramirez	40	30.7896	9.2104	84.831
Rich Niles	40	42.6316	–2.6316	6.925
Mike Kiel	50	42.6316	7.3684	54.293
Mark Reynolds	30	42.6316	–12.6316	159.557
Soni Jones	70	54.4736	15.5264	241.069
			0.0000	784.211

The standard error of estimate is 9.901, found by using formula (13–6).

$$s_{y \cdot x} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}} = \sqrt{\frac{784.211}{10 - 2}} = 9.901$$

The deviations (Y – \hat{Y}) are the vertical deviations from the regression line. To illustrate, the 10 deviations from Table 13–4 are shown in Chart 13–13. Note in Table 13–4 that the sum of the signed deviations is zero. This indicates that the positive deviations (above the regression line) are offset by the negative deviations (below the regression line).

**CHART 13–13** Sales Calls and Copiers Sold for 10 Salespeople

Statistical software eases computation when you are finding the least squares equation, the standard error of estimate, and the coefficient of correlation as well as other regression statistics. A portion of the Excel output for the Copier Sales of America example is included below. The intercept and slope values are cells F13 and F14, the standard error of estimate is in F8, and the coefficient of correlation (called Multiple R) is in cell F5.



Sales Representative	Calls	Sales
Tom Miller	20	32
Jeff Hall	10	62
Ellen Vincent	20	42
Ging Rich	30	62
Susan Wilson	10	32
Carlos Ramirez	10	42
Trish Niles	10	42
Mike Lee	10	52
Mark Reynolds	20	32
Don Jones	30	72

SUMMARY OUTPUT	
Regression Statistics	
R Square	0.769114102
R Square (Adjusted)	0.576102412
Adjusted R Square	0.5211622
Standard Error	14.88177685
Observations	
Coefficients	
Intercept	18.94738642
Calls	1.184210522

Thus far we have presented linear regression only as a descriptive tool. In other words it is a simple summary ($\hat{Y} = a + bX$) of the relationship between the dependent Y variable and the independent X variable. When our data are a sample taken from a population, we are doing inferential statistics. Then we need to recall the distinction between population parameters and sample statistics. In this case, we “model” the linear relationship in the population by the equation:

$$Y = \alpha + \beta X$$

where

Y is any value of the dependent variable.

α is the Y -intercept (the value of Y when $X = 0$) in the population.

β is the slope (the amount by which Y changes when X increases by one unit) of the population line.

X is any value of the independent variable.

Now α and β are population parameters and a and b , respectively, are estimates of those parameters. They are computed from a particular sample taken from the population. Fortunately, the formulas given earlier in the chapter for a and b do not change when we move from using regression as a descriptive tool to regression in statistical inference.

It should be noted that the linear regression equation for the sample of salespeople is only an estimate of the relationship between the two variables for the population. Thus, the values of a and b in the regression equation are usually referred to as the **estimated regression coefficients**, or simply the **regression coefficients**.

Assumptions Underlying Linear Regression

To properly apply linear regression, several assumptions are necessary. Chart 13–14 illustrates these assumptions.

1. For each value of X , there are corresponding Y values. These Y values follow the normal distribution.
2. The means of these normal distributions lie on the regression line.
3. The standard deviations of these normal distributions are all the same. The best estimate we have of this common standard deviation is the standard error of estimate ($s_{y \cdot x}$).

Statistics in Action

Studies indicate that for both men and women, those who are considered good looking earn higher wages than those who are not. In addition, for men there is a correlation between height and salary. For each additional inch of height, a man can expect to earn an additional \$250 per year. So a man 6'6" tall receives a \$3,000 “stature” bonus over his 5'6" counterpart. Being overweight or underweight is also related to earnings, particularly among women. A study of young women showed the heaviest 10 percent earned about 6 percent less than their lighter counterparts.

Linear Regression and Correlation

481

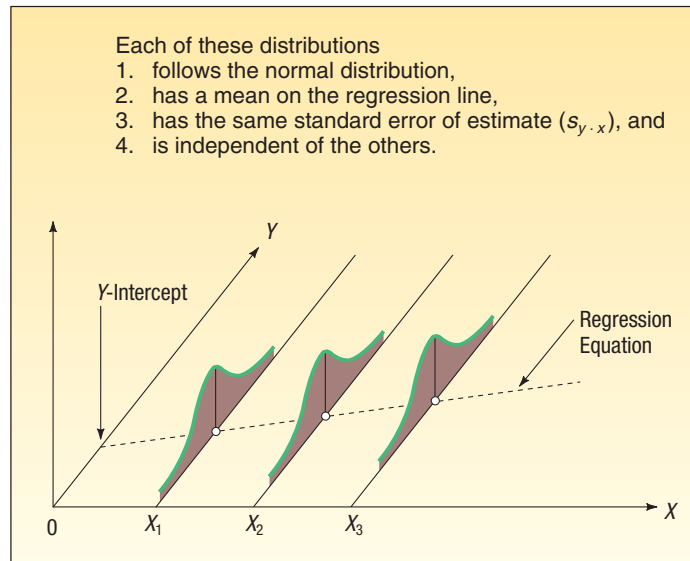


CHART 13–14 Regression Assumptions Shown Graphically

4. The Y values are statistically independent. This means that in selecting a sample a particular X does not depend on any other value of X . This assumption is particularly important when data are collected over a period of time. In such situations, the errors for a particular time period are often correlated with those of other time periods.

Recall from Chapter 7 that if the values follow a normal distribution, then the mean plus or minus one standard deviation will encompass 68 percent of the observations, the mean plus or minus two standard deviations will encompass 95 percent of the observations, and the mean plus or minus three standard deviations will encompass virtually all of the observations. The same relationship exists between the predicted values \hat{Y} and the standard error of estimate ($s_{y \cdot x}$).

1. $\hat{Y} \pm s_{y \cdot x}$ will include the middle 68 percent of the observations.
2. $\hat{Y} \pm 2s_{y \cdot x}$ will include the middle 95 percent of the observations.
3. $\hat{Y} \pm 3s_{y \cdot x}$ will include virtually all the observations.

We can now relate these assumptions to Copier Sales of America, where we studied the relationship between the number of sales calls and the number of copiers sold. Assume that we took a much larger sample than $n = 10$, but that the standard error of estimate was still 9.901. If we drew a parallel line 9.901 units above the regression line and another 9.901 units below the regression line, about 68 percent of the points would fall between the two lines. Similarly, a line 19.802 [$2s_{y \cdot x} = 2(9.901)$] units above the regression line and another 19.802 units below the regression line should include about 95 percent of the data values.

As a rough check, refer to the second column from the right in Table 13–4 on page 479, i.e., the column headed “Deviation.” Three of the 10 deviations exceed one standard error of estimate. That is, the deviation of -12.6316 for Tom Keller, -12.6316 for Mark Reynolds, and $+15.5264$ for Soni Jones all exceed the value of 9.901, which is one standard error from the regression line. All of the values are within 19.802 units of the regression line. To put it another way, 7 of the 10 deviations in the sample are within one standard error of the regression line and all are within two—a good result for a relatively small sample.

Self-Review 13–4



Refer to Self-Reviews 13–1 and 13–3, where the owner of Haverty's Furniture was studying the relationship between sales and the amount spent on advertising. Determine the standard error of estimate.

Exercises

21. Refer to Exercise 13.
 - a. Determine the standard error of estimate.
 - b. Suppose a large sample is selected (instead of just five). About 68 percent of the predictions would be between what two values?
22. Refer to Exercise 14.
 - a. Determine the standard error of estimate.
 - b. Suppose a large sample is selected (instead of just eight). About 95 percent of the predictions would be between what two values?
23. Refer to Exercise 15.
 - a. Determine the standard error of estimate.
 - b. Suppose a large sample is selected (instead of just 10). About 95 percent of the predictions regarding kilowatt-hours would occur between what two values?
24. Refer to Exercise 16.
 - a. Determine the standard error of estimate.
 - b. Suppose a large sample is selected (instead of just 10). About 95 percent of the predictions regarding sales would occur between what two values?
25. Refer to Exercise 5. Determine the standard error of estimate.
26. Refer to Exercise 6. Determine the standard error of estimate.

Confidence and Prediction Intervals

The standard error of estimate is also used to establish confidence intervals when the sample size is large and the scatter around the regression line approximates the normal distribution. In our example involving the number of sales calls and the number of copiers sold, the sample size is small; hence, we need a correction factor to account for the size of the sample. In addition, when we move away from the mean of the independent variable, our estimates are subject to more variation, and we also need to adjust for this.

We are interested in providing interval estimates of two types. The first, which is called a **confidence interval**, reports the *mean* value of Y for a given X . The second type of estimate is called a **prediction interval**, and it reports the *range of values* of Y for a *particular* value of X . To explain further, suppose we estimate the salary of executives in the retail industry based on their years of experience. If we want an interval estimate of the mean salary of *all* retail executives with 20 years of experience, we calculate a confidence interval. If we want an estimate of the salary of Curtis Bender, a *particular* retail executive with 20 years of experience, we calculate a prediction interval.

To determine the confidence interval for the mean value of Y for a given X , the formula is:

**CONFIDENCE INTERVAL
FOR THE MEAN OF Y ,
GIVEN X**

$$\hat{Y} \pm t(s_{y \cdot x}) \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X - \bar{X})^2}}$$

[13–7]

Linear Regression and Correlation

483

where

 \hat{Y} is the predicted value for any selected X value. X is any selected value of X . \bar{X} is the mean of the X s, found by $\Sigma X/n$. n is the number of observations. $s_{y \cdot x}$ is the standard error of estimate. t is the value of t from Appendix B.2 with $n - 2$ degrees of freedom.

We first described the t distribution in Chapter 9. In review the concept of t was developed by William Gossett in the early 1900s. He noticed that $\bar{X} \pm z s_{\bar{X}}$ was not precisely correct for small samples. He observed, for example, for degrees of freedom of 120, that 95 percent of the items fell within $\bar{X} \pm 1.98s$ instead of $\bar{X} \pm 1.96s_{\bar{X}}$. This difference is not too critical, but note what happens as the sample size becomes smaller:

<i>df</i>	<i>t</i>
120	1.980
60	2.000
21	2.080
10	2.228
3	3.182

This is logical. The smaller the sample size, the larger the possible error. The increase in the t value compensates for this possibility.

Example

We return to the Copier Sales of America illustration. Determine a 95 percent confidence interval for all sales representatives who make 25 calls and a prediction interval for Sheila Baker, a West Coast sales representative who made 25 calls.

Solution

We use formula (13–7) to determine a confidence interval. Table 13–5 includes the necessary totals and a repeat of the information of Table 13–2 on page 462.

TABLE 13–5 Calculations Needed for Determining the Confidence Interval and Prediction Interval

Sales Representative	Sales Calls, (X)	Copier Sales, (Y)	$(X - \bar{X})$	$(X - \bar{X})^2$
Tom Keller	20	30	–2	4
Jeff Hall	40	60	18	324
Brian Virost	20	40	–2	4
Greg Fish	30	60	8	64
Susan Welch	10	30	–12	144
Carlos Ramirez	10	40	–12	144
Rich Niles	20	40	–2	4
Mike Kiel	20	50	–2	4
Mark Reynolds	20	30	–2	4
Soni Jones	30	70	8	64
			0	760

The first step is to determine the number of copiers we expect a sales representative to sell if he or she makes 25 calls. It is 48.5526, found by $\hat{Y} = 18.9476 + 1.1842X = 18.9476 + 1.1842(25)$.

To find the t value, we need to first know the number of degrees of freedom. In this case the degrees of freedom is $n - 2 = 10 - 2 = 8$. We set the confidence level at 95 percent. To find the value of t , move down the left-hand column of Appendix B.2 to 8 degrees of freedom, then move across to the column with the 95 percent level of confidence. The value of t is 2.306.

In the previous section we calculated the standard error of estimate to be 9.901. We let $X = 25$, $\bar{X} = \Sigma X/n = 220/10 = 22$, and from Table 13–5 $\Sigma(X - \bar{X})^2 = 760$. Inserting these values in formula (13–7), we can determine the confidence interval.

$$\begin{aligned}\text{Confidence Interval} &= \hat{Y} \pm t_{s_{y \cdot x}} \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \\ &= 48.5526 \pm 2.306(9.901) \sqrt{\frac{1}{10} + \frac{(25 - 22)^2}{760}} \\ &= 48.5526 \pm 7.6356\end{aligned}$$

Thus, the 95 percent confidence interval for all sales representatives who make 25 calls is from 40.9170 up to 56.1882. To interpret, let's round the values. If a sales representative makes 25 calls, he or she can expect to sell 48.6 copiers. It is likely those sales will range from 40.9 to 56.2 copiers.

To determine the prediction interval for a particular value of Y for a given X , formula (13–7) is modified slightly: A 1 is added under the radical. The formula becomes:

**PREDICTION INTERVAL
FOR Y , GIVEN X**

$$\hat{Y} \pm t_{s_{y \cdot x}} \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \quad [13-8]$$

Suppose we want to estimate the number of copiers sold by Sheila Baker, who made 25 sales calls. The 95 percent prediction interval is determined as follows:

$$\begin{aligned}\text{Prediction Interval} &= \hat{Y} \pm t_{s_{y \cdot x}} \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \\ &= 48.5526 \pm 2.306(9.901) \sqrt{1 + \frac{1}{10} + \frac{(25 - 22)^2}{760}} \\ &= 48.5526 \pm 24.0746\end{aligned}$$

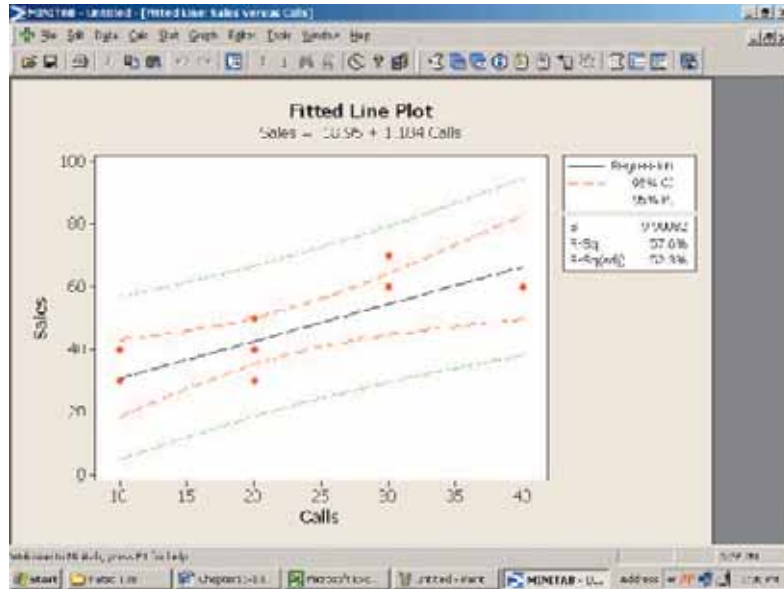
Thus, the interval is from 24.478 up to 72.627 copiers. We conclude that the number of copiers sold will be between about 24 and 73 for a particular sales representative who makes 25 calls. This interval is quite large. It is much larger than the confidence interval for all sales representatives who made 25 calls. It is logical, however, that there should be more variation in the sales estimate for an individual than for a group.

The following MINITAB graph shows the relationship between the regression line (in the center), the confidence interval (shown in crimson), and the prediction interval (shown in green). The bands for the prediction interval are always further from the regression line than those for the confidence interval. Also, as the values of X move away from the mean number of calls (22) in either the positive or the negative direction, the confidence interval and prediction interval bands widen. This is caused by the numerator of the right-hand term under the radical in formulas (13–7) and (13–8). That is, as the term $(X - \bar{X})^2$ increases, the widths of the confidence interval and the prediction interval also increase. To put it another way, there is less

Linear Regression and Correlation

485

precision in our estimates as we move away, in either direction, from the mean of the independent variable.



We wish to emphasize again the distinction between a confidence interval and a prediction interval. A confidence interval refers to all cases with a given value of X and is computed by formula (13–7). A prediction interval refers to a particular case for a given value of X and is computed using formula (13–8). The prediction interval will always be wider because of the extra 1 under the radical in the second equation.

Self-Review 13–5



Refer to the sample data in Self-Reviews 13–1, 13–3, and 13–4, where the owner of Haverty's Furniture was studying the relationship between sales and the amount spent on advertising. The sales information for the last four months is repeated below.

Month	Advertising Expense (\$ million)	Sales Revenue (\$ million)
July	2	7
August	1	3
September	3	8
October	4	10

The regression equation was computed to be $\hat{Y} = 1.5 + 2.2X$, and the standard error 0.9487. Both variables are reported in millions of dollars. Determine the 90 percent confidence interval for the typical month in which \$3 million was spent on advertising.

Exercises

27. Refer to Exercise 13.
 - a. Determine the .95 confidence interval for the mean predicted when $X = 7$.
 - b. Determine the .95 prediction interval for an individual predicted when $X = 7$.
28. Refer to Exercise 14.
 - a. Determine the .95 confidence interval for the mean predicted when $X = 7$.
 - b. Determine the .95 prediction interval for an individual predicted when $X = 7$.

29. Refer to Exercise 15.
- Determine the .95 confidence interval, in thousands of kilowatt-hours, for the mean of all six-room homes.
 - Determine the .95 prediction interval, in thousands of kilowatt-hours, for a particular six-room home.
30. Refer to Exercise 16.
- Determine the .95 confidence interval, in thousands of dollars, for the mean of all sales personnel who make 40 contacts.
 - Determine the .95 prediction interval, in thousands of dollars, for a particular salesperson who makes 40 contacts.

More on the Coefficient of Determination

Earlier in the chapter, on page 465, we defined the coefficient of determination as the percent of the variation in the dependent variable that is accounted for by the independent variable. We indicated it is the square of the coefficient of correlation and that it is written r^2 .

To further examine the basic concept of the coefficient of determination, suppose there is interest in the relationship between years on the job, X , and weekly production, Y . Sample data revealed:

Employee	Years on Job, X	Weekly Production, Y
Gordon	14	6
James	7	5
Ford	3	3
Salter	15	9
Artes	11	7

The sample data were plotted in a scatter diagram. Since the relationship between X and Y appeared to be linear, a line was drawn through the plots (see Chart 13–15). The equation is $\hat{Y} = 2 + 0.4X$.

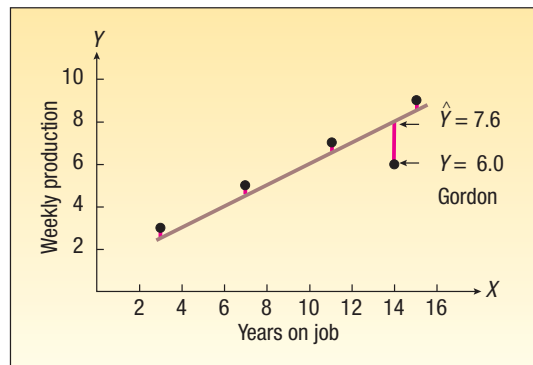


CHART 13–15 Observed Data and the Least Squares Line

Note in Chart 13–15 that, if we were to use that line to predict weekly production for an employee, in no case would our prediction be exact. That is, there would be some error in each of our predictions. As an example, for Gordon, who has been with the company 14 years, we would predict weekly production to be 7.6 units; however, he produces only 6 units.

Linear Regression and Correlation

487

Unexplained variation

To measure the overall error in our prediction, every deviation from the line is squared and the squares summed. The predicted point on the line is designated \hat{Y} , read Y hat, and the observed point is designated Y . For Gordon, $(Y - \hat{Y})^2 = (6 - 7.6)^2 = (-1.6)^2 = 2.56$. Logically, this variation cannot be explained by the independent variable, so it is referred to as the *unexplained variation*. Specifically, we cannot explain why Gordon's production of 6 units is 1.6 units below his predicted production of 7.6 units, based on the number of years he has been on the job.

The sum of the squared deviations, $\Sigma(Y - \hat{Y})^2$, is 4.00. (See Table 13–6.) The term $\Sigma(Y - \hat{Y})^2 = 4.00$ is the variation in Y (production) that cannot be predicted from X . It is the “unexplained” variation in Y .

TABLE 13–6 Computations Needed for the Unexplained Variation

	X	Y	\hat{Y}	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
Gordon	14	6	7.6	−1.6	2.56
James	7	5	4.8	0.2	0.04
Ford	3	3	3.2	−0.2	0.04
Salter	15	9	8.0	1.0	1.00
Artes	11	7	6.4	0.6	0.36
Total	$\overline{50}$	$\overline{30}$		$\overline{0.0^*}$	$\overline{4.00}$

*Must be 0.

Total variation in Y

Now suppose *only* the Y values (weekly production, in this problem) are known and we want to predict production for every employee. The actual production figures for the employees are 6, 5, 3, 9, and 7 (from Table 13–6). To make these predictions, we could assign the mean weekly production (6 units, found by $\Sigma Y/n = 30/5 = 6$) to each employee. This would keep the sum of the squared prediction errors at a minimum. (Recall from Chapter 3 that the sum of the squared deviations from the arithmetic mean for a set of numbers is smaller than the sum of the squared deviations from any other value, such as the median.) Table 13–7 shows the necessary calculations. The sum of the squared deviations is 20, as shown in Table 13–7. The value 20 is referred to as the *total variation in Y* .

TABLE 13–7 Calculations Needed for the Total Variation in Y

Name	Weekly Production, Y	Mean Weekly Production, \bar{Y}	$Y - \bar{Y}$	$(Y - \bar{Y})^2$
Gordon	6	6	0	0
James	5	6	−1	1
Ford	3	6	−3	9
Salter	9	6	3	9
Artes	7	6	1	1
Total			$\overline{0^*}$	$\overline{20}$

*Must be 0.

What we did to arrive at the total variation in Y is shown diagrammatically in Chart 13–16.

Logically, the total variation in Y can be subdivided into unexplained variation and explained variation. To arrive at the explained variation, since we know the total variation and unexplained variation, we simply subtract: Explained variation = Total variation − Unexplained variation. Dividing the explained variation by the total

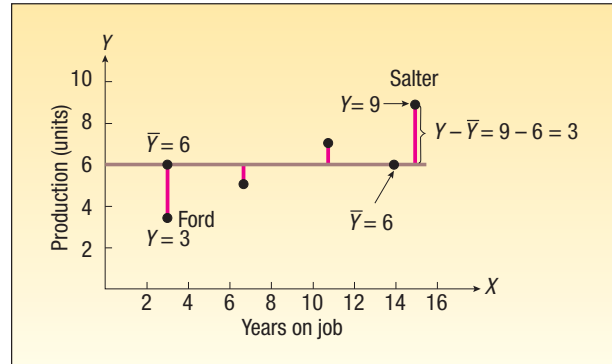


CHART 13–16 Plots Showing Deviations from the Mean of Y

variation gives the coefficient of determination, r^2 , which is a proportion. In terms of a formula:

**COEFFICIENT OF
DETERMINATION**

$$r^2 = \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}}$$

$$= \frac{\sum(Y - \bar{Y})^2 - \sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2}$$

[13–9]

In this problem:

$$r^2 = \frac{20 - 4}{20} = \frac{16}{20}$$

Table 13–7
Table 13–6
Explained variation
Total variation

$$= .80$$

As mentioned, .80 is a proportion. We say that 80 percent of the variation in weekly production, Y , is determined, or accounted for, by its linear relationship with X (years on the job).

As a check, formula (13–1) for the coefficient of correlation could be used. Squaring r gives the coefficient of determination, r^2 . Exercise 31 offers a check on the preceding problem.

Exercises

31. Using the preceding problem, involving years on the job and weekly production, verify that the coefficient of determination is in fact .80.
32. The number of shares of Icom, Inc., turned over during a month, and the price at the end of the month, are listed in the following table. Also given are the \hat{Y} values.

Turnover (thousands of shares), X	Actual Price, Y	Estimated Price, \hat{Y}
4	\$2	\$2.7
1	1	0.6
5	4	3.4
3	2	2.0
2	1	1.3

Linear Regression and Correlation

489

- a. Draw a scatter diagram. Plot a line through the dots.
- b. Compute the coefficient of determination using formula (13–10).
- c. Interpret the coefficient of determination.

The Relationships among the Coefficient of Correlation, the Coefficient of Determination, and the Standard Error of Estimate

In an earlier section, we discussed the standard error of estimate, which measures how close the actual values are to the regression line. When the standard error is small, it indicates that the two variables are closely related. In the calculation of the standard error, the key term is $\Sigma(Y - \hat{Y})^2$. If the value of this term is small, then the standard error will also be small.

The correlation coefficient measures the strength of the linear association between two variables. When the points on the scatter diagram appear close to the line, we note that the correlation coefficient tends to be large. Thus, the standard error of estimate and the coefficient of correlation relate the same information but use a different scale to report the strength of the association. However, both measures involve the term $\Sigma(Y - \hat{Y})^2$.

We also noted that the square of the correlation coefficient is the coefficient of determination. The coefficient of determination measures the percent of the variation in Y that is explained by the variation in X .

A convenient vehicle for showing the relationship among these three measures is an ANOVA table. This table is similar to the analysis of variance table developed in Chapter 12. In that chapter, the total variation was divided into two components: that due to the *treatments* and that due to *random error*. The concept is similar in regression analysis. The total variation, $\Sigma(Y - \bar{Y})^2$, is divided into two components: (1) that explained by the *regression* (explained by the independent variable) and (2) the *error*, or unexplained variation. These two categories are identified in the first column of the ANOVA table that follows. The column headed “*df*” refers to the degrees of freedom associated with each category. The total number of degrees of freedom is $n - 1$. The number of degrees of freedom in the regression is 1, since there is only one independent variable. The number of degrees of freedom associated with the error term is $n - 2$. The term “SS” located in the middle of the ANOVA table refers to the sum of squares—the variation. The terms are computed as follows:

$$\text{Regression} = \text{SSR} = \Sigma(\hat{Y} - \bar{Y})^2$$

$$\text{Error variation} = \text{SSE} = \Sigma(Y - \hat{Y})^2$$

$$\text{Total variation} = \text{SS total} = \Sigma(Y - \bar{Y})^2$$

The format for the ANOVA table is:

Source	<i>df</i>	SS	MS
Regression	1	SSR	SSR/1
Error	$n - 2$	SSE	SSE/($n - 2$)
Total	$n - 1$	SS total*	

*SS total = SSR + SSE.

The coefficient of determination, r^2 , can be obtained directly from the ANOVA table by:

$$\text{COEFFICIENT OF DETERMINATION} \quad r^2 = \frac{\text{SSR}}{\text{SS total}} = 1 - \frac{\text{SSE}}{\text{SS total}} \quad [13-10]$$

The term “SSR/SS total” is the proportion of the variation in Y explained by the independent variable, X . Note the effect of the SSE term on r^2 . As SSE decreases, r^2 will increase. To put it another way, as the standard error decreases, the r^2 term increases.

The standard error of estimate can also be obtained from the ANOVA table using the following equation:

$$\text{STANDARD ERROR OF ESTIMATE} \quad s_{y \cdot x} = \sqrt{\frac{\text{SSE}}{n - 2}} \quad [13-11]$$

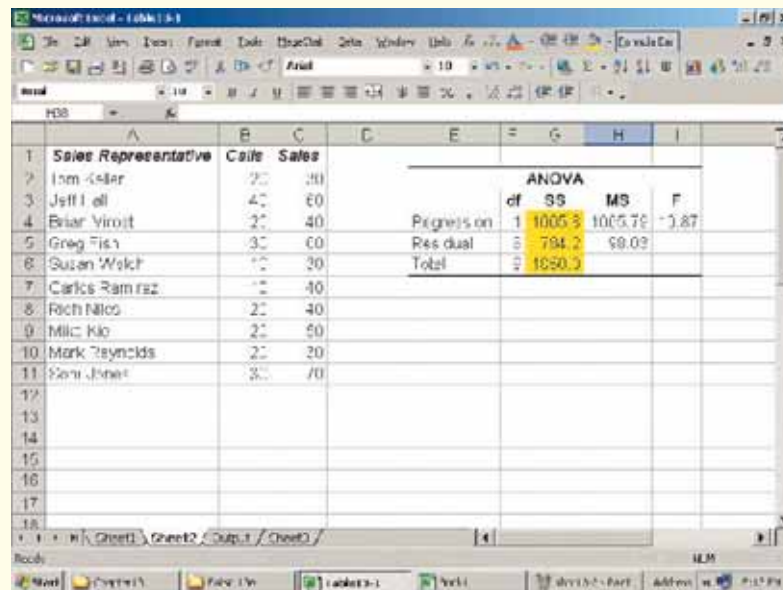
The Copier Sales of America example is used to illustrate the computations of the coefficient of determination and the standard error of estimate from an ANOVA table.

Example

In the Copier Sales of America example we studied the relationship between the number of sales calls made and the number of copiers sold. Use a computer software package to determine the least squares regression equation and the ANOVA table. Identify the regression equation, the standard error of estimate, and the coefficient of determination on the computer output. From the ANOVA table on the computer output, determine the coefficient of determination and the standard error of estimate using formulas (13-10) and (13-11).

Solution

The output from Excel follows.



Sales Representative	Calls	Sales
Tom Keller	20	30
Jeff Hall	40	60
Belar Virolit	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Milt Klo	20	60
Mark Reynolds	20	20
Sam Jones	30	70

ANOVA				
	df	SS	MS	F
Regression	1	1065.8	1065.76	1.87
Residual	8	794.2	99.28	
Total	9	1860.0		

From formula (13-10) the coefficient of determination is .576, found by

$$r^2 = \frac{\text{SSR}}{\text{SS total}} = \frac{1,065.8}{1,860} = .576$$

This is the same value we computed earlier in the chapter, when we found the coefficient of determination by squaring the coefficient of correlation. Again, the interpre-



Linear Regression and Correlation

491

tation is that the independent variable, Calls, explains 57.6 percent of the variation in the number of copiers sold. If we needed the coefficient of correlation, we could find it by taking the square root of the coefficient of determination:

$$r = \sqrt{r^2} = \sqrt{.576} = .759$$

A problem does remain, and that involves the sign for the coefficient of correlation. Recall that the square root of a value could have either a positive or a negative sign. The sign of the coefficient of correlation will always be the same as that of the slope. That is, b and r will always have the same sign. In this case the sign of the regression coefficient (b) is positive, so the coefficient of correlation is .759.

To find the standard error of estimate, we use formula (13–11):

$$s_{y \cdot x} = \sqrt{\frac{SSE}{n - 2}} = \sqrt{\frac{784.2}{10 - 2}} = 9.901$$

Again, this is the same value calculated earlier in the chapter. These values are identified on the Excel computer output.

Transforming Data



The coefficient of correlation describes the strength of the *linear* relationship between two variables. It could be that two variables are closely related, but their relationship is not linear. Be cautious when you are interpreting the coefficient of correlation. A value of r may indicate there is no linear relationship, but it could be there is a relationship of some other nonlinear or curvilinear form.

To explain, below is a listing of 22 professional golfers, the number of events in which they participated, the amount of their winnings, and their mean score for the 2004 season. In golf, the objective is to play 18 holes in the least number of strokes. So, we would expect that those golfers with the lower

mean scores would have the larger winnings. In other words, score and winnings should be inversely related.

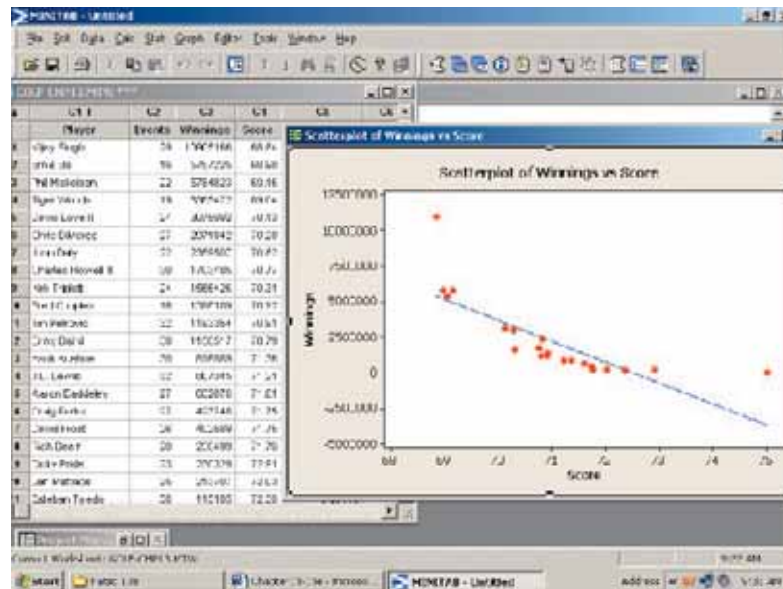
In 2004 Tiger Woods played in 19 events, earned \$5,365,472, and had a mean score per round of 69.04. Fred Couples played in 16 events, earned \$1,396,109, and had a mean score per round of 70.92. The data for the 22 golfers follows.

Player	Events	Winnings	Score
Vijay Singh	29	\$10,905,166	68.84
Ernie Els	16	5,787,225	68.98
Phil Mickelson	22	5,784,823	69.16
Tiger Woods	19	5,365,472	69.04
Davis Love III	24	3,075,092	70.13
Chris DiMarco	27	2,971,842	70.28
John Daly	22	2,359,507	70.82
Charles Howell III	30	1,703,485	70.77
Kirk Triplett	24	1,566,426	70.31
Fred Couples	16	1,396,109	70.92
Tim Petrovic	32	1,193,354	70.91

continued

Player	Events	Winnings	Score
Briny Baird	30	\$1,156,517	70.79
Hank Kuehne	30	816,889	71.36
J. L. Lewis	32	807,345	71.21
Aaron Baddeley	27	632,876	71.61
Craig Perks	27	423,748	71.75
David Frost	26	402,589	71.75
Rich Beem	28	230,499	71.76
Dicky Pride	23	230,329	72.91
Len Mattiace	25	213,707	72.03
Esteban Toledo	36	115,185	72.36
David Gossett	25	21,250	75.01

The correlation between the variables Winnings and Score is -0.782 . This is a fairly strong inverse relationship. However, when we plot the data on a scatter diagram the relationship does not appear to be linear; it does not seem to follow a straight line. See the scatter diagram on the right-hand side of the following MINITAB output. The data points for the lowest score and the highest score seem to be well away from the regression line. In addition, for the scores between 70 and 72, the winnings are below the regression line. If the relationship were linear, we would expect these points to be both above and below the line.



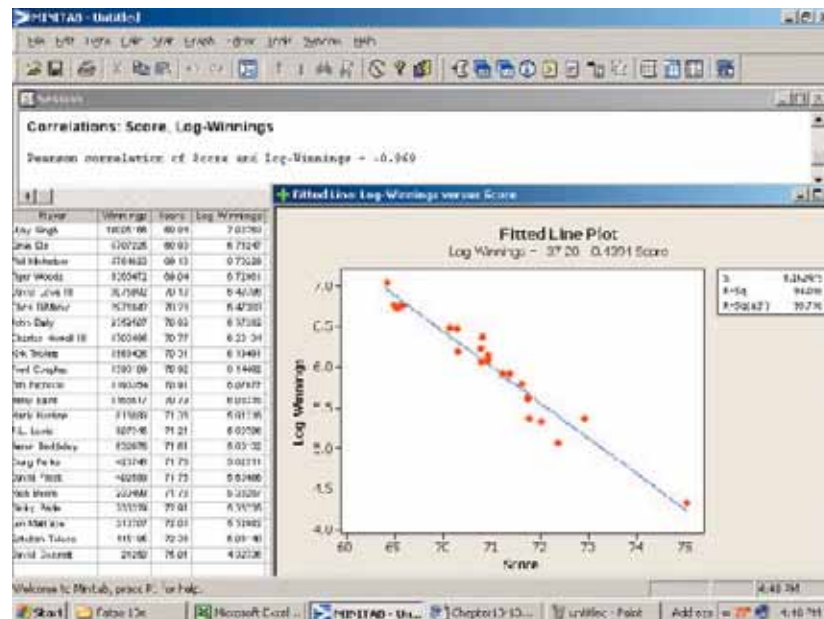
What can we do to explore other (nonlinear) relationships? One possibility is to transform one of the variables. For example, instead of using Y as the dependent variable, we might use its log, reciprocal, square, or square root. Another possibility is to transform the independent variable in the same way. There are other transformations, but these are the most common.

In the golf winnings example, changing the scale of the dependent variable is effective. We determine the log of each golfer's winnings and then find the correlation between the log of winnings and score. That is, we find the log to the base 10 of Tiger Woods' earnings of \$5,365,472, which is 6.72961. Next we find the log to the base 10 of each golfer's winnings and then determine the correlation between log of winnings and the score. The correlation coefficient increases from -0.782 to -0.969 . This means that the coefficient of determination is .939 [$r^2 = (-0.969)^2 = .939$]. That is, 93.9 percent of the variation in the log of winnings is accounted for by the independent variable score.

Linear Regression and Correlation

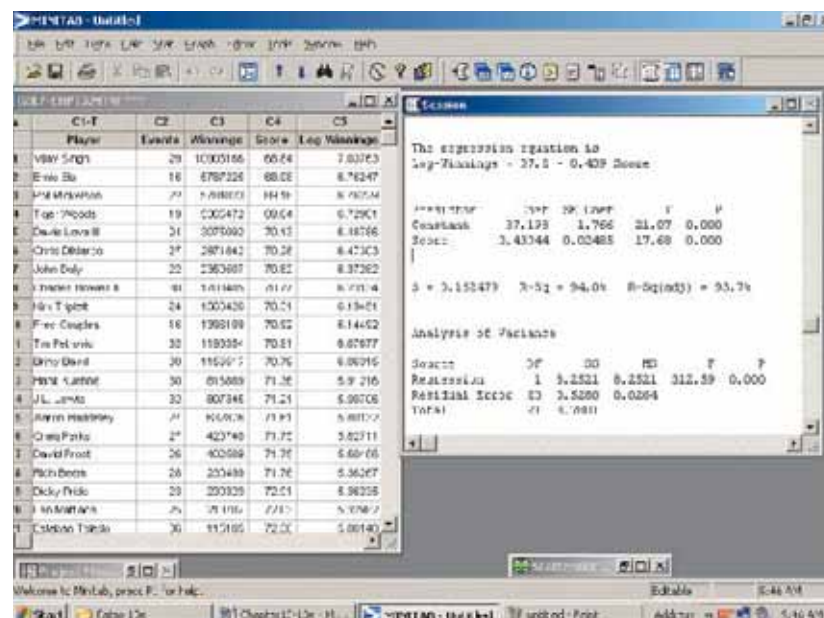
493

We have determined an equation that fits the data more closely than the straight line did. Clearly, as the mean score for a golfer increases, he can expect his winnings to decrease. It no longer appears that some of the data points are different from the regression line, as we found when using winnings instead of the log of winnings as the dependent variable. Also note the points between 70 and 72 in particular are now randomly distributed above and below the regression line.



We can also estimate the amount of winnings based on the score. Following is the MINITAB regression output using score as the independent variable and the log of winnings as the dependent variable. Based on the regression equation, a golfer with a mean score of 70 could expect to earn:

$$\hat{Y} = 37.198 - .43944X = 37.198 - .43944(70) = 6.4372$$



The value 6.4372 is the log to the base 10 of winnings. The antilog of 6.4372 is 2,736,528. So a golfer that had a mean score of 70 could expect to earn \$2,736,528. We can also evaluate the change in scores. The above golfer had a mean score of 70 and estimated earnings of \$2,736,528. How much less would a golfer expect to win if his mean score was 71? Again solving the regression equation:

$$\hat{Y} = 37.198 - .43944X = 37.198 - .43944(71) = 5.99776$$

The antilog of this value is \$994,855. So based on the regression analysis, there is a large financial incentive for a professional golfer to reduce his mean score by even one stroke. Those of you that play golf or know a golfer understand how difficult that change would be! That one stroke is worth over \$1,700,000.

Exercises

33. Given the following ANOVA table:

SOURCE	DF	SS	MS	F
Regression	1	1000.0	1000.00	26.00
Error	13	500.0	38.46	
Total	14	1500.0		

- Determine the coefficient of determination.
 - Assuming a direct relationship between the variables, what is the coefficient of correlation?
 - Determine the standard error of estimate.
34. On the first statistics exam the coefficient of determination between the hours studied and the grade earned was 80 percent. The standard error of estimate was 10. There were 20 students in the class. Develop an ANOVA table.
35. Given the following sample observations, develop a scatter diagram. Compute the coefficient of correlation. Does the relationship between the variables appear to be linear? Try squaring the X-variable and then determine the correlation coefficient.

X	−8	−16	12	2	18
Y	58	247	153	3	341

36. According to basic economics, as the demand for a product increases, the price will decrease. Listed below is the number of units demanded and the price.

Demand	Price
2	\$120.0
5	90.0
8	80.0
12	70.0
16	50.0
21	45.0
27	31.0
35	30.0
45	25.0
60	21.0

- Determine the correlation between price and demand. Plot the data in a scatter diagram. Does the relationship seem to be linear?
- Transform the price to a log to the base 10. Plot the log of the price and the demand. Determine the correlation coefficient. Does this seem to improve the relationship between the variables?

Covariance (Optional)

To understand the coefficient of correlation let's begin by plotting data. Chart 13–4 on page 463 is a scatter diagram of Copier Sales of America data. Observe that as the number of sales calls increases, so does the number of copiers sold. The

Linear Regression and Correlation

495

number of units sold is scaled on the vertical axis and the number of sales calls on the horizontal axis.

Let's compute again the mean of both the sales calls (X) and the number of units sold (Y). From Table 13–2 on page 462, the mean number of sales calls is 22.0, found by $220/10$. The mean number of units sold is 45, found by $450/10$. So we conclude a typical sales representative for Copier Sales of America makes 22 sales calls and sells 45 copiers in a month. In Chart 13–4, we have moved the origin from the point $(0, 0)$ to the points (\bar{X}, \bar{Y}) . This will allow us to understand the association between the number of sales calls and the number of copiers sold.

At this point we can make some interpretations of the data. As discussed earlier, if the points are scattered in all four quadrants, then it is likely there is little association between the variables. A predominance of data points in the lower-left and upper-right quadrants indicates a positive relationship, while data points in the upper-left and lower-right quadrants suggest a negative relationship.

To evaluate the relationship you noted visually in Chart 13–4, compute the term $\Sigma(X - \bar{X})(Y - \bar{Y})$. Notice the thrust of this term. It is the sum of the products of the deviations between the number of sales calls and the mean number of sales calls and the number of copiers sold and the mean number of copiers sold, for each of the 10 sales representatives. For a point located in the upper-right quadrant (Quadrant I) both the X and Y values would be larger than their means. From Table 13–2 Soni Jones made 30 sales calls and sold 70 copiers. Both of the values are larger than the mean of 22 sales calls and 45 copiers sold. The product of these deviations $(30 - 22)(70 - 45) = 200$. Other points in this quadrant will also have a positive result.

Points located in the upper-left quadrant (Quadrant IV) will have a negative value. Mike Kiel, for example, made 20 sales calls and sold 50 copiers. So $(X - \bar{X})(Y - \bar{Y}) = (20 - 22)(50 - 45) = -10$.

So the value from points in Quadrant IV will offset (deduct from) those in Quadrant I. If the term $\Sigma(X - \bar{X})(Y - \bar{Y})$ is a positive value, this indicates a positive relationship between the two variables. A negative value indicates a negative relationship between the variables. The symbol SS_{xy} is used to identify this term. It is computed from the following formula.

$$SS_{xy} = \Sigma(X - \bar{X})(Y - \bar{Y})$$

The term SS_{xy} found by using the above formula, indicates the relationship between the X and Y variables. However, it is difficult to interpret because (a) the units involved will be mixed between those of X and Y , and (b) the term could be made larger by merely increasing the sample size. To control for the sample size, the term is divided by $n - 1$, the sample size minus 1. This is the same procedure we used in determining the sample variance, described in Chapter 3. The result is called the **covariance**.

SAMPLE COVARIANCE

$$s_{xy} = \frac{SS_{xy}}{n - 1}$$

[13–12]

Returning to our Copier Sales of America problem, the covariance is 100. See the details in the Excel printout.

$$s_{xy} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{n - 1} = \frac{900}{10 - 1} = 100$$



Sales Representative	Calls	Sales	(X - \bar{X})	(Y - \bar{Y})	(X - \bar{X})(Y - \bar{Y})
Sam Miller	20.00	30.00	-2	-10	20
Jeff All	40.00	40.00	8	0	0
Chris Vandy	20.00	30.00	-2	-5	10
Greg Fish	20.00	40.00	-2	10	-20
Susan Welch	10.00	30.00	-12	-5	60
George Harston	30.00	40.00	2	0	0
Rich Pines	20.00	40.00	-2	10	-20
Mike Hill	20.00	30.00	-2	-5	10
Mark Reynolds	30.00	30.00	2	-5	-10
Sam Jones	20.00	40.00	-2	10	-20
Mean	22.00	35.00			
Standard deviation	9.19	14.34			
Sum					900

How do we interpret the covariance? Recall that the variance summarizes the variability of a single variable. The covariance summarizes the relationship *between* two variables. It differs from the variance in that it can assume negative values. A negative covariance indicates that the two variables are inversely related. The covariance is difficult to interpret, because of the units involved. In this case does a covariance of 100 indicate the variables are closely related or that they are not related at all? We cannot tell. We can only conclude that because this is a positive value, the two variables are positively related. A second difficulty involves the units of the two variables. In this example, one variable is in number of calls and the other units sold. So the units of the results are not familiar.

To remove the problem with the units, the covariance is standardized. That is, it is divided by the standard deviations of X and Y . The result is the coefficient of correlation.

We can verify the coefficient of correlation in the Copier Sales of America example starting on page 462. The first step is to compute the standard deviation of the number of sales calls and the number of copiers sold. Using the data in Table 13–2 the standard deviations are:

$$s_y = \sqrt{\frac{1,850}{10 - 1}} = 14.34$$

$$s_x = \sqrt{\frac{760}{10 - 1}} = 9.19$$

The term SS_{xy} is 900, found by

$$SS_{xy} = \sum (X - \bar{X})(Y - \bar{Y}) = 900$$

The covariance s_{xy} is found by

$$s_{xy} = \frac{SS_{xy}}{n - 1} = \frac{(900)}{9} = 100.0$$

Finally, the correlation is 0.759, the same as that determined using Formula 13–1 on page 464.

$$r = \frac{s_{xy}}{s_x s_y} = \frac{100.0}{(9.19)(14.34)} = 0.759$$

Linear Regression and Correlation

497

Exercises

- Write a brief description of the coefficient of correlation. What is its range of values? What does it mean when it is zero? Under what conditions can it be larger than 1.00?
- Describe what is meant by the covariance. Can it be negative? What is its range of values?
- A phone company executive is studying the relationship between the number of telephone calls per week and the number of people in the household. A sample of 12 families is obtained.

Calls (Y)	22	15	20	31	75	26	20	28	26	59	23	33
Family (X)	4	5	4	3	7	5	6	5	5	7	2	5

Plot the information in a scatter diagram. Compute the covariance and the coefficient of correlation. Is the relationship direct or inverse, strong or weak?

- The director of the Tampa Zoo is studying the relationship between the number of admissions, in thousands, and the high temperature, in degrees Fahrenheit. A sample of 15 days is selected and the sample information is reported below.

Admissions (000s)	Temperature (°F)	Admissions (000s)	Temperature (°F)
2.0	86	2.2	84
0.6	71	2.5	66
2.0	89	1.3	76
2.1	73	3.6	84
2.2	76	1.0	75
2.1	75	1.8	72
0.5	68	2.1	76
0.3	72		

Plot the information in a scatter diagram. Compute the covariance and the coefficient of correlation. Is the relationship direct or inverse? Would you consider the association strong or weak?

Chapter Summary

- A scatter diagram is a graphic tool to portray the relationship between two variables.
 - The dependent variable is scaled on the Y-axis and is the variable being estimated.
 - The independent variable is scaled on the X-axis and is the variable used as the estimator.
- The coefficient of correlation measures the strength of the linear association between two variables.
 - Both variables must be at least the interval scale of measurement.
 - The coefficient of correlation can range from -1.00 up to 1.00 .
 - If the correlation between two variables is 0 , there is no association between them.
 - A value of 1.00 indicates perfect positive correlation, and -1.00 perfect negative correlation.
 - A positive sign means there is a direct relationship between the variables, and a negative sign means there is an inverse relationship.
 - It is designated by the letter r and found by the following equation:

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{(n - 1)s_x s_y} \quad [13-1]$$

- The following equation is used to determine whether the correlation in the population is different from 0 .

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{with } n-2 \text{ degrees of freedom} \quad [13-2]$$

- III. The coefficient of determination is the fraction of the variation in one variable that is explained by the variation in the other variable.
- It ranges from 0 to 1.0.
 - It is the square of the coefficient of correlation.
- IV. In regression analysis we estimate one variable based on another variable.
- The variable being estimated is the dependent variable.
 - The variable used to make the estimate is the independent variable.
 - The relationship between the variables must be linear.
 - Both the independent and the dependent variables must be interval or ratio scale.
 - The least squares criterion is used to determine the regression equation.
- V. The least squares regression line is of the form $\hat{Y} = a + bX$.
- \hat{Y} is the estimated value of Y for a selected value of X .
 - a is the constant or intercept.
 - It is the value of \hat{Y} when $X = 0$.
 - a is computed using the following equation.

$$a = \bar{Y} - b\bar{X} \quad [13-5]$$

- b is the slope of the fitted line.
 - It shows the amount of change in \hat{Y} for a change of one unit in X .
 - A positive value for b indicates a direct relationship between the two variables, and a negative value an inverse relationship.
 - The sign of b and the sign of r , the coefficient of correlation, are always the same.
 - b is computed using the following equation.

$$b = r \left(\frac{s_y}{s_x} \right) \quad [13-4]$$

- X is the value of the independent variable.
- VI. The standard error of estimate measures the variation around the regression line.
- It is in the same units as the dependent variable.
 - It is based on squared deviations from the regression line.
 - Small values indicate that the points cluster closely about the regression line.
 - It is computed using the following formula.

$$s_{y \cdot x} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - 2}} \quad [13-6]$$

- VII. Inference about linear regression is based on the following assumptions.
- For a given value of X , the values of Y are normally distributed about the line of regression.
 - The standard deviation of each of the normal distributions is the same for all values of X and is estimated by the standard error of estimate.
 - The deviations from the regression line are independent, with no pattern to the size or direction.
- VIII. There are two types of interval estimates.
- In a confidence interval the mean value of Y is estimated for a given value of X .
 - It is computed from the following formula.

$$\hat{Y} \pm t(s_{y \cdot x}) \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X - \bar{X})^2}} \quad [13-7]$$

- The width of the interval is affected by the level of confidence, the size of the standard error of estimate, and the size of the sample, as well as the value of the independent variable.
- In a prediction interval the individual value of Y is estimated for a given value of X .
 - It is computed from the following formula.

$$\hat{Y} \pm t_{s_{y \cdot x}} \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X - \bar{X})^2}} \quad [13-8]$$

- The difference between formulas (13-7) and (13-8) is the 1 under the radical.
 - The prediction interval will be wider than the confidence interval.
 - The prediction interval is also based on the level of confidence, the size of the standard error of estimate, the size of the sample, and the value of the independent variable.

Linear Regression and Correlation

499

Pronunciation Key

SYMBOL	MEANING	PRONUNCIATION
ΣXY	Sum of the products of X and Y	<i>Sum X Y</i>
ρ	Coefficient of correlation in the population	<i>Rho</i>
\hat{Y}	Estimated value of Y	<i>Y hat</i>
$s_{y \cdot x}$	Standard error of estimate	<i>s sub y dot x</i>
r^2	Coefficient of determination	<i>r square</i>

Chapter Exercises

37. A regional commuter airline selected a random sample of 25 flights and found that the correlation between the number of passengers and the total weight, in pounds, of luggage stored in the luggage compartment is 0.94. Using the .05 significance level, can we conclude that there is a positive association between the two variables?
38. A sociologist claims that the success of students in college (measured by their GPA) is related to their family's income. For a sample of 20 students, the coefficient of correlation is 0.40. Using the 0.01 significance level, can we conclude that there is a positive correlation between the variables?
39. An Environmental Protection Agency study of 12 automobiles revealed a correlation of 0.47 between engine size and emissions. At the .01 significance level, can we conclude that there is a positive association between these variables? What is the p -value? Interpret.
40. A suburban hotel derives its gross income from its hotel and restaurant operations. The owners are interested in the relationship between the number of rooms occupied on a nightly basis and the revenue per day in the restaurant. Below is a sample of 25 days (Monday through Thursday) from last year showing the restaurant income and number of rooms occupied.

Day	Income	Occupied	Day	Income	Occupied
1	\$1,452	23	14	\$1,425	27
2	1,361	47	15	1,445	34
3	1,426	21	16	1,439	15
4	1,470	39	17	1,348	19
5	1,456	37	18	1,450	38
6	1,430	29	19	1,431	44
7	1,354	23	20	1,446	47
8	1,442	44	21	1,485	43
9	1,394	45	22	1,405	38
10	1,459	16	23	1,461	51
11	1,399	30	24	1,490	61
12	1,458	42	25	1,426	39
13	1,537	54			

Use a statistical software package to answer the following questions.

- a. Does the breakfast revenue seem to increase as the number of occupied rooms increases? Draw a scatter diagram to support your conclusion.
 - b. Determine the coefficient of correlation between the two variables. Interpret the value.
 - c. Is it reasonable to conclude that there is a positive relationship between revenue and occupied rooms? Use the .10 significance level.
 - d. What percent of the variation in revenue in the restaurant is accounted for by the number of rooms occupied?
41. The table below shows the number of cars (in millions) sold in the United States for various years and the percent of those cars manufactured by GM.

Year	Cars Sold (millions)	Percent GM	Year	Cars Sold (millions)	Percent GM
1950	6.0	50.2	1980	11.5	44.0
1955	7.8	50.4	1985	15.4	40.1
1960	7.3	44.0	1990	13.5	36.0
1965	10.3	49.9	1995	15.5	31.7
1970	10.1	39.5	2000	17.4	28.6
1975	10.8	43.1	2003	17.1	27.8

Use a statistical software package to answer the following questions.

- Is the number of cars sold directly or indirectly related to GM's percent of the market? Draw a scatter diagram to show your conclusion.
 - Determine the coefficient of correlation between the two variables. Interpret the value.
 - Is it reasonable to conclude that there is a negative association between the two variables? Use the .01 significance level.
 - How much of the variation in GM's market share is accounted for by the variation in cars sold?
42. For a sample of 32 large U.S. cities, the correlation between the mean number of square feet per office worker and the mean monthly rental rate in the central business district is $-.363$. At the .05 significance level, can we conclude that there is a negative association in the population between the two variables?
43. What is the relationship between the amount spent per week on recreation and the size of the family? Do larger families spend more on recreation? A sample of 10 families in the Chicago area revealed the following figures for family size and the amount spent on recreation per week.

Family Size	Amount Spent on Recreation	Family Size	Amount Spent on Recreation
3	\$ 99	3	\$111
6	104	4	74
5	151	4	91
6	129	5	119
6	142	3	91

- Compute the coefficient of correlation.
 - Determine the coefficient of determination.
 - Can we conclude that there is a positive association between the amount spent on recreation and family size? Use the .05 significance level.
44. A sample of 12 homes sold last week in St. Paul, Minnesota, is selected. Can we conclude that, as the size of the home (reported below in thousands of square feet) increases, the selling price (reported in \$ thousands) also increases?

Home Size (thousands of square feet)	Selling Price (\$ thousands)	Home Size (thousands of square feet)	Selling Price (\$ thousands)
1.4	100	1.3	110
1.3	110	0.8	85
1.2	105	1.2	105
1.1	120	0.9	75
1.4	80	1.1	70
1.0	105	1.1	95

- Compute the coefficient of correlation.
- Determine the coefficient of determination.
- Can we conclude that there is a positive association between the size of the home and the selling price? Use the .05 significance level.

Linear Regression and Correlation

501

45. The manufacturer of Cardio Glide exercise equipment wants to study the relationship between the number of months since the glide was purchased and the length of time the equipment was used last week.

Person	Months Owned	Hours Exercised	Person	Months Owned	Hours Exercised
Rupple	12	4	Massa	2	8
Hall	2	10	Sass	8	3
Bennett	6	8	Karl	4	8
Longnecker	9	5	Malrooney	10	2
Phillips	7	5	Veights	5	5

- Plot the information on a scatter diagram. Let hours of exercise be the dependent variable. Comment on the graph.
 - Determine the coefficient of correlation. Interpret.
 - At the .01 significance level, can we conclude that there is a negative association between the variables?
46. The following regression equation was computed from a sample of 20 observations:

$$\hat{Y} = 15 - 5X$$

SSE was found to be 100 and SS total 400.

- Determine the standard error of estimate.
 - Determine the coefficient of determination.
 - Determine the coefficient of correlation. (Caution: Watch the sign!)
47. An ANOVA table is:

SOURCE	DF	SS	MS	F
Regression	1	50		
Error				
Total	24	500		

- Complete the ANOVA table.
 - How large was the sample?
 - Determine the standard error of estimate.
 - Determine the coefficient of determination.
48. Following is a regression equation.

$$\hat{Y} = 17.08 + 0.16X$$

This information is also available: $s_{y \cdot x} = 4.05$, $\Sigma(X - \bar{X})^2 = 1,030$, and $n = 5$.

- Estimate the value of \hat{Y} when $X = 50$.
 - Develop a 95 percent prediction interval for an individual value of Y for $X = 50$.
49. The National Highway Association is studying the relationship between the number of bidders on a highway project and the winning (lowest) bid for the project. Of particular interest is whether the number of bidders increases or decreases the amount of the winning bid.

Project	Number of Bidders, X	Winning Bid (\$ millions), Y	Project	Number of Bidders, X	Winning Bid (\$ millions), Y
1	9	5.1	9	6	10.3
2	9	8.0	10	6	8.0
3	3	9.7	11	4	8.8
4	10	7.8	12	7	9.4
5	5	7.7	13	7	8.6
6	10	5.5	14	7	8.1
7	7	8.3	15	6	7.8
8	11	5.5			

- a. Determine the regression equation. Interpret the equation. Do more bidders tend to increase or decrease the amount of the winning bid?
- b. Estimate the amount of the winning bid if there were seven bidders.
- c. A new entrance is to be constructed on the Ohio Turnpike. There are seven bidders on the project. Develop a 95 percent prediction interval for the winning bid.
- d. Determine the coefficient of determination. Interpret its value.
50. Mr. William Profit is studying companies going public for the first time. He is particularly interested in the relationship between the size of the offering and the price per share. A sample of 15 companies that recently went public revealed the following information.

Company	Size (\$ millions), <i>X</i>	Price per Share, <i>Y</i>	Company	Size (\$ millions), <i>X</i>	Price per Share, <i>Y</i>
1	9.0	10.8	9	160.7	11.3
2	94.4	11.3	10	96.5	10.6
3	27.3	11.2	11	83.0	10.5
4	179.2	11.1	12	23.5	10.3
5	71.9	11.1	13	58.7	10.7
6	97.9	11.2	14	93.8	11.0
7	93.5	11.0	15	34.4	10.8
8	70.0	10.7			

- a. Determine the regression equation.
- b. Determine the coefficient of determination. Do you think Mr. Profit should be satisfied with using the size of the offering as the independent variable?
51. Bardi Trucking Co., located in Cleveland, Ohio, makes deliveries in the Great Lakes region, the Southeast, and the Northeast. Jim Bardi, the president, is studying the relationship between the distance a shipment must travel and the length of time, in days, it takes the shipment to arrive at its destination. To investigate, Mr. Bardi selected a random sample of 20 shipments made last month. Shipping distance is the independent variable, and shipping time is the dependent variable. The results are as follows:

Shipment	Distance (miles)	Shipping Time (days)	Shipment	Distance (miles)	Shipping Time (days)
1	656	5	11	862	7
2	853	14	12	679	5
3	646	6	13	835	13
4	783	11	14	607	3
5	610	8	15	665	8
6	841	10	16	647	7
7	785	9	17	685	10
8	639	9	18	720	8
9	762	10	19	652	6
10	762	9	20	828	10

- a. Draw a scatter diagram. Based on these data, does it appear that there is a relationship between how many miles a shipment has to go and the time it takes to arrive at its destination?
- b. Determine the coefficient of correlation. Can we conclude that there is a positive correlation between distance and time? Use the .05 significance level.
- c. Determine and interpret the coefficient of determination.
- d. Determine the standard error of estimate.
52. Super Markets, Inc., is considering expanding into the Scottsdale, Arizona, area. You as director of planning, must present an analysis of the proposed expansion to the operating committee of the board of directors. As a part of your proposal, you need to include information on the amount people in the region spend per month for grocery items. You would also like to include information on the relationship between the amount spent for grocery items and income. Your assistant gathered the following sample information. The data is available on the data disk supplied with the text.

Linear Regression and Correlation

503

Household	Amount Spent	Monthly Income
1	\$ 555	\$4,388
2	489	4,558
⋮	⋮	⋮
39	1,206	9,862
40	1,145	9,883

- Let the amount spent be the dependent variable and monthly income the independent variable. Create a scatter diagram, using a software package.
 - Determine the regression equation. Interpret the slope value.
 - Determine the coefficient of correlation. Can you conclude that it is greater than 0?
53. Below is information on the price per share and the dividend for a sample of 30 companies. The sample data are available on the data disk supplied with the text.

Company	Price per Share	Dividend
1	\$20.00	\$ 3.14
2	22.01	3.36
⋮	⋮	⋮
29	77.91	17.65
30	80.00	17.36

- Calculate the regression equation using selling price based on the annual dividend. Interpret the slope value.
 - Determine the coefficient of determination. Interpret its value.
 - Determine the coefficient of correlation. Can you conclude that it is greater than 0 using the .05 significance level?
54. A highway employee performed a regression analysis of the relationship between the number of construction work-zone fatalities and the number of unemployed people in a state. The regression equation is $\text{Fatalities} = 12.7 + 0.000114 (\text{Unemp})$. Some additional output is:

Predictor	Coef	SE Coef	T	P	
Constant	12.726	8.115	1.57	0.134	
Unemp	0.00011386	0.00002896	3.93	0.001	
Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	10354	10354	15.46	0.001
Residual Error	18	12054	670		
Total	19	22408			

- How many states were in the sample?
 - Determine the standard error of estimate.
 - Determine the coefficient of determination.
 - Determine the coefficient of correlation.
 - At the .05 significance level does the evidence suggest there is a positive association between fatalities and the number unemployed?
55. A regression analysis relating the current market value in dollars to the size in square feet of homes in Greene County, Tennessee follows. The regression equation is: $\text{Value} = -37,186 + 65.0 \text{ Size}$.

Predictor	Coef	SE Coef	T	P	
Constant	-37186	4629	-8.03	0.000	
Size	64.993	3.047	21.33	0.000	
Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	13548662082	13548662082	454.98	0.000
Residual Error	33	982687392	29778406		
Total	34	14531349474			

- a. How many homes were in the sample?
 - b. Compute the standard error of estimate.
 - c. Compute the coefficient of determination.
 - d. Compute the coefficient of correlation.
 - e. At the .05 significance level does the evidence suggest a positive association between the market value of homes and the size of the home in square feet?
56. The following table shows the mean annual percent return on capital (profitability) and the mean annual percentage sales growth for eight aerospace and defense companies.

Company	Profitability	Growth
Alliant Techsystems	23.1	8.0
Boeing	13.2	15.6
General Dynamics	24.2	31.2
Honeywell	11.1	2.5
L-3 Communications	10.1	35.4
Northrop Grumman	10.8	6.0
Rockwell Collins	27.3	8.7
United Technologies	20.1	3.2

- a. Compute the coefficient of correlation. Conduct a test of hypothesis to determine if it is reasonable to conclude that the population correlation is greater than zero. Use the .05 significance level.
 - b. Develop the regression equation for profitability based on growth. Comment on the slope value.
 - c. Use a software package to determine the residual for each observation. Which company has the largest residual?
57. The following data show the retail price for 12 randomly selected laptop computers along with their corresponding processor speeds in gigahertz.

Computers	Speed	Price	Computers	Speed	Price
1	2.0	\$2,689	7	2.0	\$2,929
2	1.6	1,229	8	1.6	1,849
3	1.6	1,419	9	2.0	2,819
4	1.8	2,589	10	1.6	2,669
5	2.0	2,849	11	1.0	1,249
6	1.2	1,349	12	1.4	1,159

- a. Develop a linear equation that can be used to describe how the price depends on the processor speed.
 - b. Based on your regression equation, is there one machine that seems particularly over- or underpriced?
 - c. Compute the correlation coefficient between the two variables. At the .05 significance level conduct a test of hypothesis to determine if the population correlation could be greater than zero.
58. A consumer buying cooperative tested the effective heating area of 20 different electric space heaters with different wattages. Here are the results.

Heater	Wattage	Area	Heater	Wattage	Area
1	1,500	205	11	1,250	116
2	750	70	12	500	72
3	1,500	199	13	500	82
4	1,250	151	14	1,500	206
5	1,250	181	15	2,000	245
6	1,250	217	16	1,500	219
7	1,000	94	17	750	63
8	2,000	298	18	1,500	200
9	1,000	135	19	1,250	151
10	1,500	211	20	500	44

Linear Regression and Correlation

505

- a. Compute the correlation between the wattage and heating area. Is there a direct or an indirect relationship?
- b. Conduct a test of hypothesis to determine if it is reasonable that the coefficient is greater than zero. Use the .05 significance level.
- c. Develop the regression equation for effective heating based on wattage.
- d. Which heater looks like the “best buy” based on the size of the residual?
59. A dog trainer is exploring the relationship between the size of the dog (weight in pounds) and its daily food consumption (measured in standard cups). Below is the result of a sample of 18 observations.

Dog	Weight	Consumption	Dog	Weight	Consumption
1	41	3	10	91	5
2	148	8	11	109	6
3	79	5	12	207	10
4	41	4	13	49	3
5	85	5	14	113	6
6	111	6	15	84	5
7	37	3	16	95	5
8	111	6	17	57	4
9	41	3	18	168	9

- a. Compute the correlation coefficient. Is it reasonable to conclude that the correlation in the population is greater than zero? Use the .05 significance level.
- b. Develop the regression equation for cups based on the dog's weight. How much does each additional cup change the estimated weight of the dog?
- c. Is one of the dogs a big undereater or overeater?
60. Waterbury Insurance Company wants to study the relationship between the amount of fire damage and the distance between the burning house and the nearest fire station. This information will be used in setting rates for insurance coverage. For a sample of 30 claims for the last year, the director of the actuarial department determined the distance from the fire station (X) and the amount of fire damage, in thousands of dollars (Y). The MegaStat output is reported below. (You can find the actual data in the data set on the CD as prb13-60.)

ANOVA table					
Source	SS	df	MS	F	
Regression	1,864.5782	1	1,864.5782	38.83	
Residual	1,344.4934	28	48.0176		
Total	3,209.0716	29			
Regression output					
Variables	Coefficients	Std. Error	t (df = 28)		
Intercept	12.3601	3.2915	3.755		
Distance-X	4.7956	0.7696	6.231		

Answer the following questions.

- a. Write out the regression equation. Is there a direct or indirect relationship between the distance from the fire station and the amount of fire damage?
- b. How much damage would you estimate for a fire 5 miles from the nearest fire station?
- c. Determine and interpret the coefficient of determination.
- d. Determine the coefficient of correlation. Interpret its value. How did you determine the sign of the correlation coefficient?
- e. Conduct a test of hypothesis to determine if there is a significant relationship between the distance from the fire station and the amount of damage. Use the .01 significance level and a two-tailed test.

61. Listed below are the movies with the largest world box office sales and their world box office budget (total amount available to spend making the picture).

Movie	Year	World Box Office (\$ million)	Adjusted Budget (\$ million)
Titanic	1997	\$1,835.00	\$ 789.30
Star Wars	1977	797.90	1,084.30
Shrek 2	2004	912.00	436.50
E.T.	1982	757.00	860.60
Star Wars: Episode I—The Phantom Menace	1999	925.50	511.70
Spider-Man	2002	806.70	419.70
LOTR: The Return of the King	2003	1,129.20	377.00
Spider-Man 2	2004	784.00	373.40
The Passion of the Christ	2004	611.80	370.30
Jurassic Park	1993	920.00	513.80
The Lord of the Rings: The Two Towers	2002	920.50	354.00
Finding Nemo	2003	853.20	339.70
Forrest Gump	1994	680.00	470.20
Harry Potter and the Sorcerer's Stone	2001	968.70	338.30
LOTR: The Fellowship of the Ring	2001	860.70	334.30
The Lion King	1994	771.90	446.20
Star Wars: Episode II—Attack of the Clones	2002	648.30	323.00
Return of the Jedi	1983	573.00	563.10
Independence Day	1996	813.10	417.50
Pirates of the Caribbean	2003	653.20	305.40
The Sixth Sense	1999	661.50	348.40
The Empire Strikes Back	1980	533.90	586.80
Home Alone	1990	533.80	401.60
The Matrix: Reloaded	2003	735.70	281.50
Meet the Fockers	2004	511.90	279.20
Shrek	2001	469.70	285.10
Harry Potter and the Chamber of Secrets	2002	866.40	272.40
The Incredibles	2004	631.20	261.40
Jaws	1975	471.00	782.70
Dr. Seuss' How the Grinch Stole Christmas	2000	340.00	290.90
Monsters, Inc.	2001	524.20	272.60
Batman	1989	413.00	375.20
Men In Black	1997	587.20	328.60
Harry Potter and the Prisoner of Azkaban	2004	789.80	249.40
Toy Story 2	1999	485.70	291.80
Bruce Almighty	2003	459.00	242.60
Raiders of the Lost Ark	1981	384.00	519.70
Twister	1996	495.00	329.70
My Big Fat Greek Wedding	2002	356.50	251.00
Ghostbusters	1984	291.60	391.70
Beverly Hills Cop	1984	316.40	416.40
Cast Away	2000	424.30	261.40
The Lost World	1997	614.40	301.00
Signs	2002	408.00	237.00
Rush Hour 2	2001	329.10	240.90
Mrs. Doubtfire	1993	423.20	315.60
Ghost	1990	517.60	306.60
Aladdin	1992	502.40	311.70
Saving Private Ryan	1998	479.30	278.10
Mission: Impossible 2	2000	545.40	241.00

Linear Regression and Correlation

507

Find the correlation between the world box office budget and world box office sales. Comment on the association between the two variables. Does it appear that the movies with large budgets result in large box office revenues?

exercises.com



62. Suppose you want to study the association between the literacy rate in a country, the population, and the country's Gross Domestic Product (GDP). Go to the website of *Information Please Almanac* (<http://www.infoplease.com>). Select the category **World & News**, and then select **Countries**. A list of 195 countries starting with Afghanistan and ending with Zimbabwe will appear. Randomly select a sample of about 20 countries. It may be convenient to use a systematic sample. In other words, randomly select 1 of the first 10 countries and then select every tenth country thereafter. Click on each country name and scan the information to find the literacy rate, the population, and the GDP. Compute the correlation among the variables. In other words, find the correlation between: literacy and population, literacy and GDP, and population and GDP. *Warning:* Be careful of the units. Sometimes population is reported in millions, other times in thousands. At the .05 significance level, can we conclude that the correlation is different from zero for each pair of variables?
63. Many real estate companies and rental agencies now publish their listings on the Web. One example is Dunes Realty Company, located in Garden City and Surfside Beaches in South Carolina. Go to the Web site <http://www.dunes.com> and select **Vacation Rentals**, then **Beach Home Search**. The indicate 5 bedrooms, accommodations for 14 people, second row (this means it is across the street from the beach), and no pool or floating dock; select a week in July or August; indicate that you are willing to spend \$8,000 per week; and then click on **Search the Beach Homes**. The output should include details on the cottages that met your criteria.
 - a. Determine the correlation between the number of baths in each cottage and the weekly rental price. Can you conclude that the correlation is greater than zero at the .05 significance level? Determine the coefficient of determination.
 - b. Determine the regression equation using the number of bathrooms as the independent variable and the price per week as the dependent variable. Interpret the regression equation.
 - c. Calculate the correlation between the number of people the cottage will accommodate and the weekly rental price. At the .05 significance level can you conclude that it is different from zero?

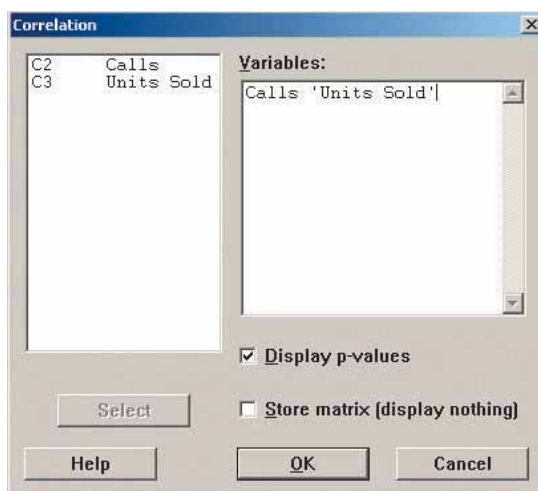
Data Set Exercises

64. Refer to the Real Estate data, which report information on homes sold in Denver, Colorado, last year.
 - a. Let selling price be the dependent variable and size of the home the independent variable. Determine the regression equation. Estimate the selling price for a home with an area of 2,200 square feet. Determine the 95 percent confidence interval and the 95 percent prediction interval for the selling price of a home with 2,200 square feet.
 - b. Let selling price be the dependent variable and distance from the center of the city the independent variable. Determine the regression equation. Estimate the selling price of a home 20 miles from the center of the city. Determine the 95 percent confidence interval and the 95 percent prediction interval for homes 20 miles from the center of the city.
 - c. Can you conclude that the independent variables "distance from the center of the city" and "selling price" are negatively correlated and that the area of the home and the selling price are positively correlated? Use the .05 significance level. Report the p -value of the test.
65. Refer to the Baseball 2005 data, which report information on the 2005 Major League Baseball season.
 - a. Let the games won be the dependent variable and total team salary, in millions of dollars, be the independent variable. Can you conclude that there is a positive

- association between the two variables? Determine the regression equation. Interpret the slope, that is, the value of b . How many additional wins will an additional \$5 million in salary bring?
- Determine the correlation between games won and ERA and between games won and team batting average. Which has the stronger correlation? Can we conclude that there is a positive correlation between wins and team batting and a negative correlation between wins and ERA? Use the .05 significance level.
 - Assume the number of games won is the dependent variable and attendance the independent variable. Can we conclude that the correlation between these two variables is greater than 0? Use the .05 significance level.
66. Refer to the Wage data, which report information on annual wages for a sample of 100 workers. Also included are variables relating to industry, years of education, and gender for each worker.
- Determine the correlation between the annual wage and the years of education. At the .05 significance level can we conclude there is a positive correlation between the two variables?
 - Determine the correlation between the annual wage and the years of work experience. At the .05 significance level can we conclude there is a positive correlation between the two variables?
67. Refer to the CIA data, which report demographic and economic information on 46 countries.
- You wish to use the labor force variable as the independent variable to predict the unemployment rate. Interpret the slope value. Use the appropriate linear regression equation to predict unemployment in the United Arab Emirates.
 - Find the correlation coefficient between the levels of exports and imports. Use the .05 significance level to test whether there is a positive correlation between these two variables.
 - Does there appear to be a relationship between the percentage of the population over 65 and the literacy percentage? Support your answer with statistical evidence. Conduct an appropriate test of hypothesis and interpret the result.

Software Commands

- The MINITAB commands for the output showing the coefficient of correlation on page 469 are:
 - Enter the sales representative's name in C1, the number of calls in C2, and the sales in C3.
 - Select **Stat**, **Basic Statistics**, and **Correlation**.
 - Select *Calls* and *Units Sold* as the variables, click on **Display p-values**, and then click **OK**.

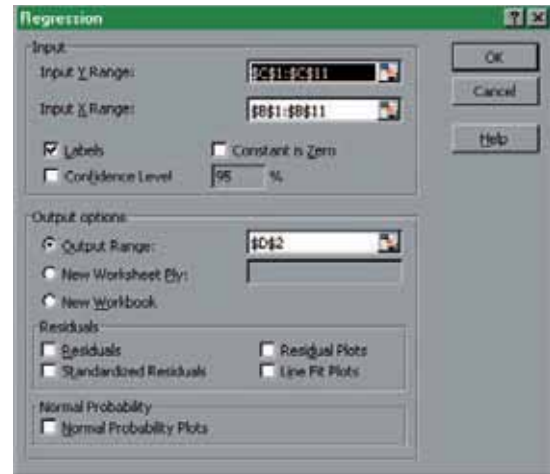


Linear Regression and Correlation

509

2. The computer commands for the Excel output on page 480 are:
 - a. Enter the variable names in row 1 of columns A, B, and C. Enter the data in rows 2 through 11 in the same columns.
 - b. Select **Tools, Data Analysis**, and then select **Regression**.
 - c. For our spreadsheet we have *Calls* in column B and *Sales* in column C. The **Input Y-Range** is *C1:C11* and the **Input X-Range** is *B1:B11*. Click on **Labels**, select *E2* as the **Output Range**, and click **OK**.

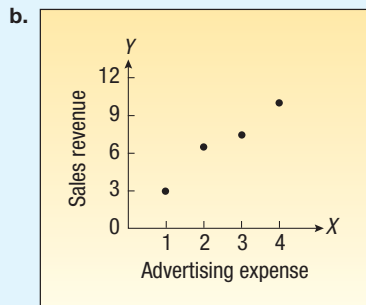
3. The MINITAB commands to the confidence intervals and prediction intervals on page 485 are:
 - a. Select **Stat, Regression**, and **Fitted line plot**.
 - b. In the next dialog box the **Response (Y)** is Sales and **Predictor (X)** is Calls. Select **Linear** for the type of regression model and then click on **Options**.
 - c. In the **Options** dialog box click on **Display confidence and prediction bands**, use the **95.0 for confidence level**, type an appropriate heading in the **Title** box, then click **OK** and then **OK** again.



Chapter 13 Answers to Self-Review



- 13–1 a.** Advertising expense is the independent variable, and sales revenue is the dependent variable.



c.

X	Y	(X - \bar{X})	(X - \bar{X}) ²	(Y - \bar{Y})	(Y - \bar{Y}) ²	(X - \bar{X})(Y - \bar{Y})
2	7	-0.5	.25	0	0	0
1	3	-1.5	2.25	-4	16	6
3	8	0.5	.25	1	1	0.5
4	10	1.5	2.25	3	9	4.5
10	28		5.00		26	11

$$\bar{X} = \frac{10}{4} = 2.5 \quad \bar{Y} = \frac{28}{4} = 7$$

$$s_x = \sqrt{\frac{5}{3}} = 1.2909944$$

$$s_y = \sqrt{\frac{26}{3}} = 2.9439203$$

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{(n - 1)s_x s_y} = \frac{11}{(4 - 1)(1.2909944)(2.9439203)} = 0.9648$$

d. There is a strong correlation between the advertising expense and sales.

e. $r^2 = .93$, 93% of the variation in sales is “explained” by variation in advertising.

- 13–2** $H_0: \rho \leq 0$, $H_1: \rho > 0$. H_0 is rejected if $t > 1.714$.

$$t = \frac{.43\sqrt{25 - 2}}{\sqrt{1 - (.43)^2}} = 2.284$$

H_0 is rejected. There is a positive correlation between the percent of the vote received and the amount spent on the campaign.

- 13–3 a.** See the calculations in Self-Review 13–1, part (c).

$$b = \frac{rs_y}{s_x} = \frac{(0.9648)(2.9439)}{1.2910} = 2.2$$

$$a = \frac{28}{4} - 2.2\left(\frac{10}{4}\right) = 7 - 5.5 = 1.5$$

b. The slope is 2.2. This indicates that an increase of \$1 million in advertising will result in an increase of \$2.2 million in sales. The intercept is 1.5. If there was no expenditure for advertising, sales would be \$1.5 million.

c. $\hat{Y} = 1.5 + 2.2(3) = 8.1$

- 13–4** 0.9487, found by:

Y	\hat{Y}	(Y - \hat{Y})	(Y - \hat{Y}) ²
7	5.9	1.1	1.21
3	3.7	-0.7	.49
8	8.1	-0.1	.01
10	10.3	-0.3	.09
			1.80

$$s_{y \cdot x} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - 2}} = \sqrt{\frac{1.80}{4 - 2}} = .9487$$

- 13–5** 6.58 and 9.62, since \hat{Y} for an X of 3 is 8.1, found by $\hat{Y} = 1.5 + 2.2(3) = 8.1$, then $\bar{X} = 2.5$ and $\sum(X - \bar{X})^2 = 5$. t from Appendix B.2 for 4 - 2 = 2 degrees of freedom at the .10 level is 2.920.

$$\begin{aligned} \hat{Y} &\pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X - \bar{X})^2}} \\ &= 8.1 \pm 2.920(0.9487)\sqrt{\frac{1}{4} + \frac{(3 - 2.5)^2}{5}} \\ &= 8.1 \pm 2.920(0.9487)(0.5477) \\ &= 6.58 \text{ and } 9.62 \text{ (in \$ millions)} \end{aligned}$$