



## BIVARIATE CORRELATION AND REGRESSION

### LEARNING OBJECTIVES

1. To comprehend the nature of correlation analysis.
2. To understand bivariate regression analysis.
3. To become aware of the coefficient of determination  $R^2$ .
4. To understand Spearman rank-order correlation.

Joe Eng and his boss, Jennifer Andersen, manage the salesforce for Big Tex Commercial Leasing in Dallas, Texas. They are in the process of doing a careful evaluation of the performance of individual salespersons in their organization. After examining the results for the past two years, Eng believes that *years of experience in the industry* is the best indicator of the volume of sales a salesperson will produce. On the other hand, Andersen believes that *number of calls made* by a salesperson is the best predictor of the sales volume. They have discussed the issue with Will Adams, the analyst who works in their department. He suggested that correlation analysis be used to resolve their debate.

Adams took the data for salespeople who have worked for the company for the last two years and ran two separate correlations. First, for each salesperson, he correlated sales with number of years in the industry. Second, he ran correlations between sales and number of calls made. The correlation coefficient between sales and years in the industry was .27. The correlation coefficient between sales and the number of calls made was .54. His conclusion is that the number of calls made is more closely associated with sales volume produced than years in the industry.

This chapter addresses correlation analysis and related techniques. After you read this chapter, you will be prepared to decide whether Adams' conclusion is correct. If it is, what are the implications for Eng and Andersen? ■



In this chapter, we will cover techniques that will permit you to evaluate the relationships between two variables.

## Bivariate Analysis of Association

➤ **bivariate techniques**  
Statistical methods of analyzing the relationship between two variables.

➤ **independent variable**  
Variable believed to affect the value of the dependent variable.

➤ **dependent variable**  
Variable expected to be explained or caused by the independent variable.

In many marketing research studies, the interests of the researcher and manager go beyond issues that can be addressed by the statistical testing of differences discussed in Chapter 15. They may be interested in the degree of association between two variables. Statistical techniques appropriate for this type of analysis are referred to as **bivariate techniques**. When more than two variables are involved, the techniques employed are known as *multivariate techniques*. Multivariate techniques are discussed in Chapter 17.

When the degree of association between two variables is analyzed, the variables are classified as the **independent** (predictor) **variable** and the **dependent** (criterion) **variable**. Independent variables are those that are believed to affect the value of the dependent variable. Independent variables such as price, advertising expenditures, or number of retail outlets may, for example, be used to predict and explain sales or market share of a brand—the dependent variable. Bivariate analysis can help provide answers to questions such as the following: How does the price of our product affect its sales? What is the relationship between household income and expenditures on entertainment?

It must be noted that none of the techniques presented in this chapter can be used to prove that one variable caused an observed change in another variable. They can be used only to describe the nature of statistical relationships between variables.

The analyst has a large number of bivariate techniques from which to choose. This chapter discusses two procedures that are appropriate for metric (ratio or internal) data—bivariate regression and Pearson's product moment correlation—and one that is appropriate for ordinal (ranking) data—Spearman rank-order correlation. Other statistical procedures that can be used for analyzing the statistical relationship between two variables include the two-group *t* test, chi-square analysis of crosstabs or contingency tables, and ANOVA (analysis of variance) for two groups. All of these procedures were introduced and discussed in Chapter 15.

## Bivariate Regression

➤ **bivariate regression analysis**  
Analysis of the strength of the linear relationship between two variables when one is considered the independent variable and the other the dependent variable.

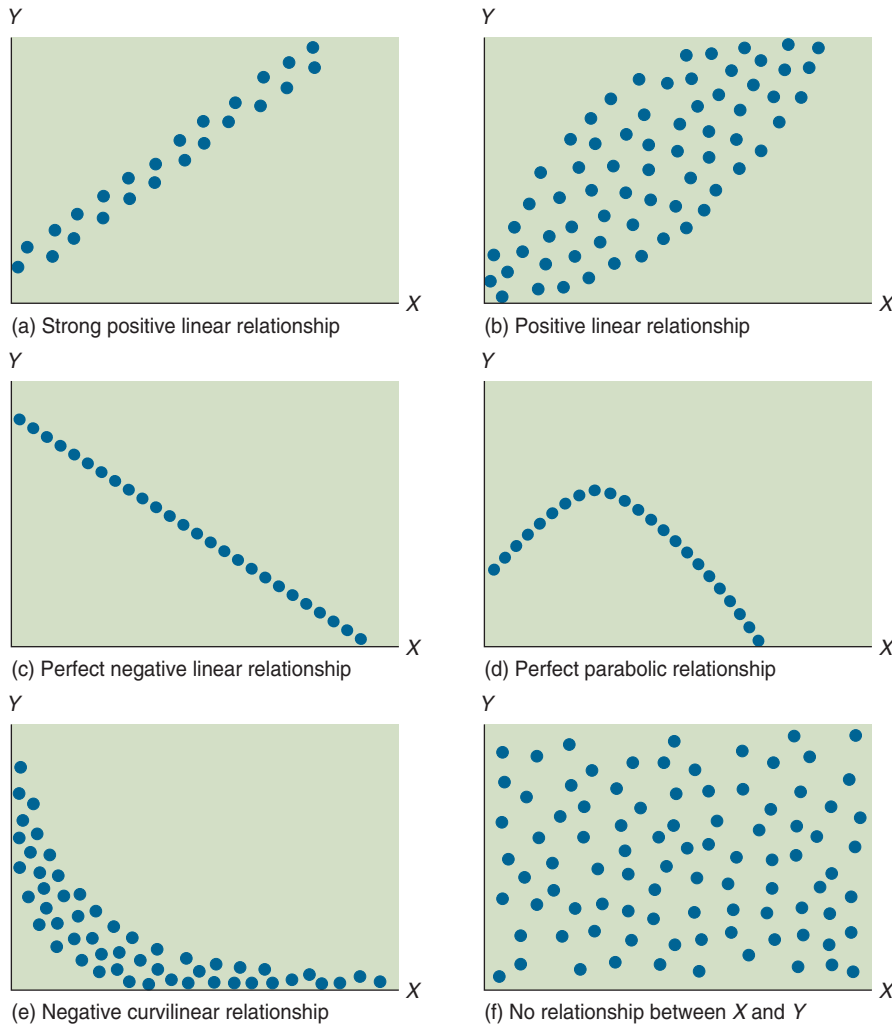
**Bivariate regression analysis** is a statistical procedure appropriate for analyzing the relationship between two variables when one is considered the dependent variable and the other the independent variable. For example, a researcher might be interested in analyzing the relationship between sales (dependent variable) and advertising (independent variable). If the relationship between advertising expenditures and sales can be accurately captured by regression analysis, the researcher can use the resulting model to predict sales for different levels of advertising. When the problem involves using two or more independent variables (for example, advertising and price) to predict the dependent variable of interest, multiple regression analysis (discussed in Chapter 17) is appropriate.

### Nature of the Relationship

One way to study the nature of the relationship between the dependent and the independent variable is to plot the data in a scatter diagram. The dependent variable *Y* is plotted on the vertical axis, whereas the independent variable *X* is plotted on the horizontal axis. By examining the scatter diagram, one can determine whether the relationship between the two variables, if any, is linear or curvilinear. If the relationship appears to be linear or close to linear, linear regression is appropriate. If a nonlinear relationship is shown in the

## Exhibit 16.1

## Types of Relationships Found in Scatter Diagrams



scatter diagram, curve-fitting nonlinear regression techniques are appropriate. These techniques are beyond the scope of this discussion.

Exhibit 16.1 depicts several kinds of underlying relationships between the  $X$  (independent) and  $Y$  (dependent) variables. Scatter diagrams (a) and (b) suggest a positive linear relationship between  $X$  and  $Y$ . However, the linear relationship shown in (b) is not as strong as that portrayed in (a); there is more scatter in the data shown in (b). Diagram (c) shows a perfect negative, or inverse, relationship between variables  $X$  and  $Y$ . An example might be the relationship between price and sales. As price goes up, sales go down. As price goes down, sales go up. Diagrams (d) and (e) show nonlinear relationships between the variables; appropriate curve-fitting techniques should be used to mathematically describe these relationships. The scatter diagram in (f) shows no relationship between  $X$  and  $Y$ .

## Example of Bivariate Regression

Stop 'N Go recently conducted a research effort designed to measure the effect of vehicular traffic past a particular store location on annual sales at that location. To control for other factors, researchers identified 20 stores that were virtually identical on all other variables known to have a significant effect on store sales (for example, square footage, amount of parking, demographics of the surrounding neighborhood). This particular

Bivariate regression analysis can help answer such questions as “How does advertising affect sales?”



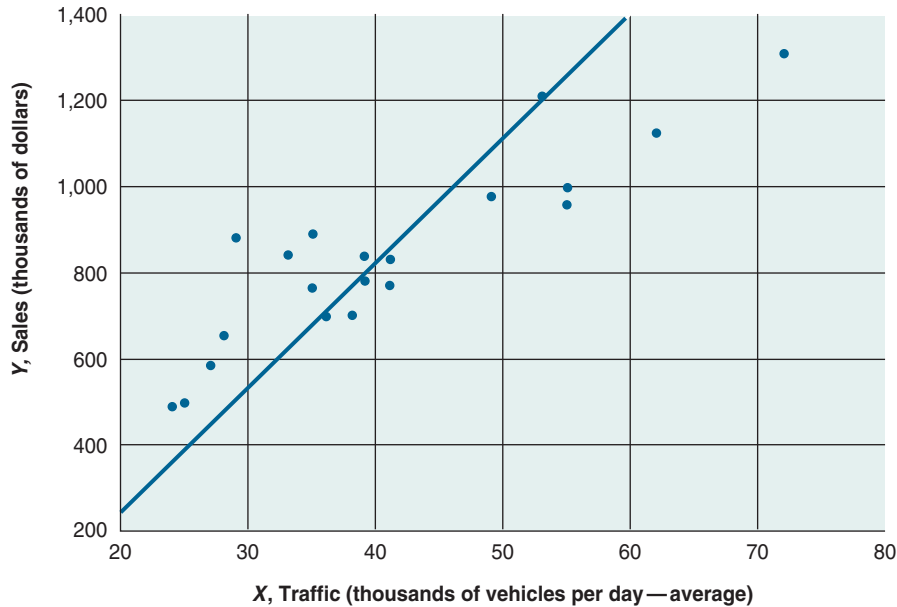
analysis is part of an overall effort by Stop 'N Go to identify and quantify the effects of various factors that affect store sales. The ultimate goal is to develop a model that can be used to screen potential sites for store locations and select, for actual purchase and store construction, the ones that will produce the highest level of sales.

After identifying the 20 sites, Stop 'N Go took a daily traffic count for each site over a 30-day period. In addition, from internal records, the company obtained total sales data for each of the 20 test stores for the preceding 12 months (see Exhibit 16.2).

EXHIBIT 16.2		Annual Sales and Average Daily Vehicular Traffic	
Store Number ( $i$ )	Average Daily Vehicular Count in Thousands ( $X_i$ )	Annual Sales in Thousands of Dollars ( $Y_i$ )	
1	62	1121	
2	35	766	
3	36	701	
4	72	1304	
5	41	832	
6	39	782	
7	49	977	
8	25	503	
9	41	773	
10	39	839	
11	35	893	
12	27	588	
13	55	957	
14	38	703	
15	24	497	
16	28	657	
17	53	1209	
18	55	997	
19	33	844	
20	29	883	

## Exhibit 16.3

## Scatterplot of Annual Sales by Traffic



A scatterplot of the resulting data is shown in Exhibit 16.3. Visual inspection of the scatterplot suggests that total sales increase as average daily vehicular traffic increases. The question now is how to characterize this relationship in a more explicit, quantitative manner.

**Least-Squares Estimation Procedure** The least-squares procedure is a fairly simple mathematical technique that can be used to fit data for  $X$  and  $Y$  to a line that best represents the relationship between the two variables. No straight line will perfectly represent every observation in the scatterplot. This is reflected in discrepancies between the actual values (dots on the scatter diagram) and predicted values (values indicated by the line). Any straight line fitted to the data in a scatterplot is subject to error. A number of lines could be drawn that would seem to fit the observations in Exhibit 16.3.

The least-squares procedure results in a straight line that fits the actual observations (dots) better than any other line that could be fitted to the observations. Put another way, the sum of the squared deviations from the line (squared differences between dots and the line) will be lower for this line than for any other line that can be fitted to the observations.

The general equation for the line is  $Y = a + bX$ . The estimating equation for regression analysis is

$$Y = \hat{a} + \hat{b}X + e$$

where

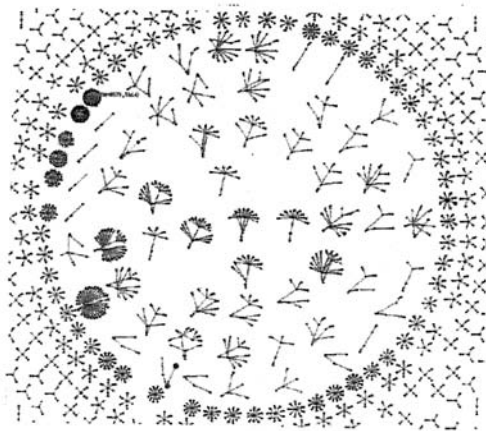
- $Y$  = dependent variable, annual sales in thousands of dollars
- $\hat{a}$  = estimated  $Y$  intercept for regression line
- $\hat{b}$  = estimated slope of regression line, regression coefficient
- $X$  = independent variable, average daily vehicular traffic in thousands of vehicles
- $e$  = error, difference between actual value and value predicted by regression line

## PRACTICING MARKETING RESEARCH

### How Data Visualization Can Help Identify Fraud

Sometimes there is so much data you cannot see the proverbial needle, or meaningful pattern, in the haystack of business information collected. Scrolling numerous pages of spreadsheets and reports, or data mining and statistical analysis, can be time-consuming and often inefficient, leaving you data-rich but information poor, comment Richard Brath and Andrea Brody. Brath is senior director of business development with Oculus Info, and Brody is president of EA Brody Consultants.

“Data visualization has the capacity to present a very large amount of detail on a single screen,” they note, and makes relationships easy to see. The approach is well-suited to identifying visual



patterns and making statistical analysis “simple and easy to comprehend.” They offer the following example from mobile phone fraud data to illustrate this point.

Thousands of phone calls have been automatically identified as potentially fraudulent. In the graph, each dot is a phone number. The dot’s size and color indicate how many calls that caller has made; the arrow points from the caller to the call recipient. Once all the data are displayed visually, we can see interconnections that were not obvious before. The special visualization technique helps you make a deeper data analysis.

First, note that around the perimeter of the diagram are clusters of small numbers of calls (typically 1–4), not too suspicious.

Second, in the center, more calls per number or caller are clustered together and may be inherently more suspicious than the ones on the perimeter because they involve a larger network of people within a larger set of possibly fraudulent calls.

Third, note at the bottom of the diagram a pair of red joined dots. The callers indicated by these large dots have phoned each other many times, suggesting that they may share a unique technique known only to themselves, and thus are worthy of suspicion.

Fourth, at the top left, note the label at the center of a small ring that indicates many calls. The caller (at the center) has called numerous people (around the ring’s edge). This caller probably is more suspicious than others because many of his or her calls are tagged as fraudulent.<sup>1</sup>

Values for  $\hat{a}$  and  $\hat{b}$  can be calculated from the following equations:

$$\hat{b} = \frac{\sum X_i Y_i - n \bar{X} \bar{Y}}{\sum X_i^2 - n(\bar{X})^2}$$

$$\hat{a} = \bar{Y} - \hat{b} \bar{X}$$

where

$\bar{X}$  = mean value of  $X$

$\bar{Y}$  = mean value of  $Y$

$n$  = sample size (number of units in the sample)

With the data from Exhibit 16.4,  $\hat{b}$  is calculated as follows:

$$\hat{b} = \frac{734,083 - 20(40.8)(841.3)}{36,526 - 20(40.8)^2} = 14.7$$

The value of  $\hat{a}$  is calculated as follows:

$$\begin{aligned} \hat{a} &= \bar{Y} - \hat{b}\bar{X} \\ &= 841.3 - 14.72(40.8) = 240.9 \end{aligned}$$

Thus, the estimated regression function is given by

$$\begin{aligned} \hat{Y} &= \hat{a} + \hat{b}X \\ &= 240.9 + 14.7(X) \end{aligned}$$

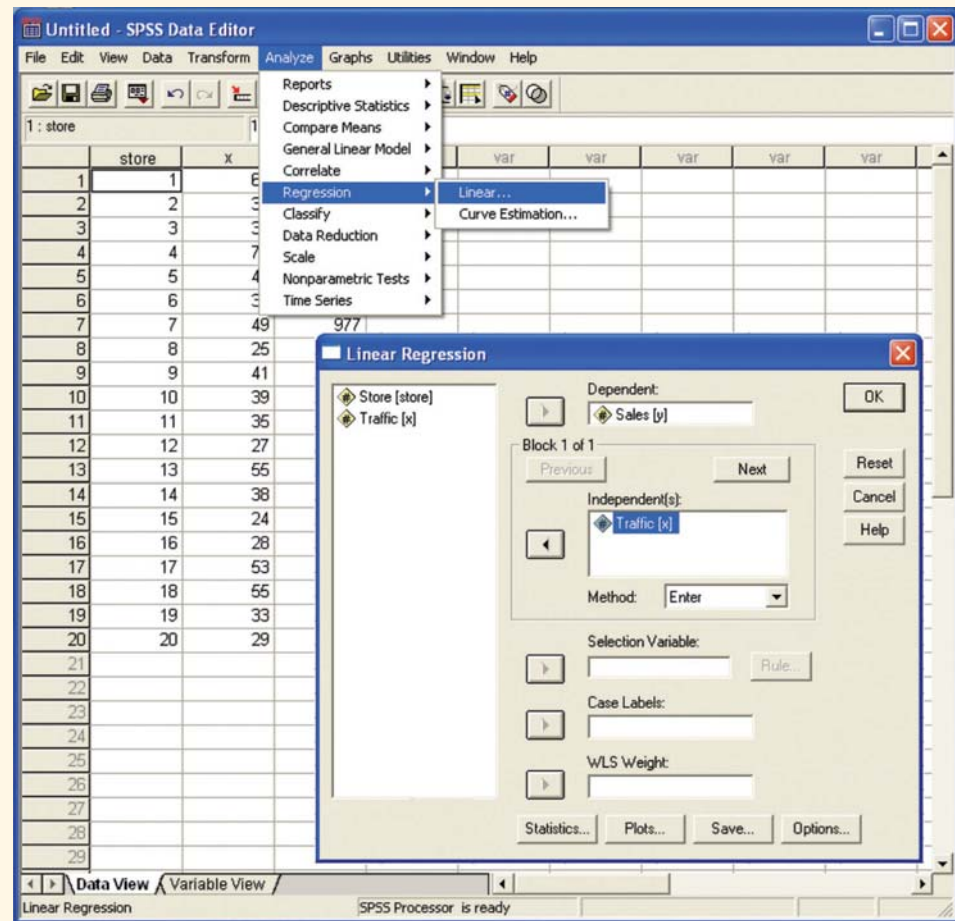
where  $\hat{Y}$  ( $Y$  hat) is the value of the estimated regression function for a given value of  $X$ .

According to the estimated regression function, for every additional 1,000 vehicles per day in traffic ( $X$ ), total annual sales will increase by \$14,720 (estimated value of  $b$ ). The value of  $\hat{a}$  is 240.9. Technically,  $\hat{a}$  is the estimated value of the dependent variable ( $Y$ , or annual sales) when the value of the independent variable ( $X$ , or average daily vehicular traffic) is zero.

EXHIBIT 16.4		Least-Squares Computation			
Store	X	Y	X <sup>2</sup>	Y <sup>2</sup>	XY
1	62	1,121	3,844	1,256,641	69,502
2	35	766	1,225	586,756	26,810
3	36	701	1,296	491,401	25,236
4	72	1,304	5,184	1,700,416	93,888
5	41	832	1,681	692,224	34,112
6	39	782	1,521	611,524	30,498
7	49	977	2,401	954,529	47,873
8	25	503	625	253,009	12,575
9	41	773	1,681	597,529	31,693
10	39	839	1,521	703,921	32,721
11	35	893	1,225	797,449	31,255
12	27	588	729	345,744	15,876
13	55	957	3,025	915,849	52,635
14	38	703	1,444	494,209	26,714
15	24	497	576	247,009	11,928
16	28	657	784	431,649	18,396
17	53	1,209	2,809	1,461,681	64,077
18	55	997	3,025	994,009	54,835
19	33	844	1,089	712,336	27,852
20	29	883	841	779,689	25,607
Sum	816	16,826	36,526	15,027,574	734,083
Mean	40.8	841.3			



Steps that you need to go through to do the bivariate regression problem shown in the book are provided below along with the output produced. Use the data set [Bivregex](#), which you can download from the Web site for the text.



## Steps in SPSS

1. Select *Analyze* → *Regression* → *Linear*.
2. Move *y* to *Dependent*.
3. Move *x* to *Independent(s)*.
4. Click *OK*.

# SPSS Output for Regression

## Regression

### Variables Entered/Removed<sup>b</sup>

Model	Variables Entered	Variables Removed	Method
1	Traffic <sup>a</sup>	.	Enter

- a. All requested variables entered.
- b. Dependent Variable: Sales

### Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.896 <sup>a</sup>	.803	.792	97.640

- a. Predictors: (Constant), Traffic

### ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	700255.40	1	700255.399	73.451	.000 <sup>a</sup>
	Residual	171604.80	18	9533.600		
	Total	871860.20	19			

- a. Predictors: (Constant), Traffic
- b. Dependent Variable: Sales

### Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	240.857	73.383		3.282	.004
	Traffic	14.717	1.717	.896	8.570	.000

- a. Dependent Variable: Sales

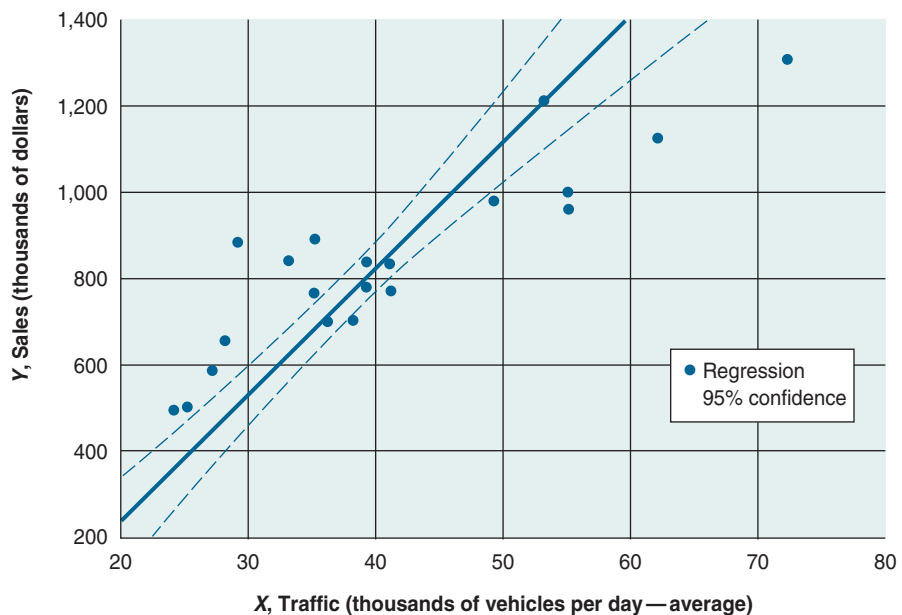
**Regression Line** Predicted values for  $Y$ , based on calculated values for  $\hat{a}$  and  $\hat{b}$ , are shown in Exhibit 16.5. In addition, errors for each observation ( $Y - \hat{Y}$ ) are shown. The regression line resulting from the  $\hat{Y}$  values is plotted in Exhibit 16.6.

**Strength of Association:  $R^2$**  The estimated regression function describes the nature of the relationship between  $X$  and  $Y$ . Another important factor is the strength of the relationship between the variables. How widely do the actual values of  $Y$  differ from the values predicted by the model?

EXHIBIT 16.5		Predicted Values and Errors for Each Observation				
Store	X	Y	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$	$(Y - \bar{Y})^2$
1	62	1,121	1,153.3	-32.2951	1,043	78,232
2	35	766	755.9	10.05716	101	5,670
3	36	701	770.7	-69.6596	4,852	19,684
4	72	1,304	1,300.5	3.537362	13	214,091
5	41	832	844.2	-12.2434	150	86
6	39	782	814.8	-32.8098	1,076	3,516
7	49	977	962.0	15.02264	226	18,414
8	25	503	608.8	-105.775	11,188	114,447
9	41	773	844.2	-71.2434	5,076	4,665
10	39	839	814.8	24.19015	585	5
11	35	893	755.9	137.0572	18,785	2,673
12	27	588	638.2	-50.2088	2,521	64,161
13	55	957	1,050.3	-93.2779	8,701	13,386
14	38	703	800.1	-97.0931	9,427	19,127
15	24	497	594.1	-97.0586	9,420	118,542
16	28	657	652.9	4.074415	17	33,966
17	53	1,209	1,020.8	188.1556	35,403	135,203
18	55	997	1,050.3	-53.2779	2,839	24,242
19	33	844	726.5	117.4907	13,804	7
20	29	883	667.6	215.3577	46,379	1,739
Sum	816	16,826	16,826		171,605	871,860
Mean	40.8	841				

**coefficient of determination**  
 Percentage of the total variation in the dependent variable explained by the independent variable.

The **coefficient of determination**, denoted by  $R^2$ , is the measure of the strength of the linear relationship between  $X$  and  $Y$ . The coefficient of determination measures the percentage of the total variation in  $Y$  that is “explained” by the variation in  $X$ . The  $R^2$  statistic ranges from 0 to 1. If there is a perfect linear relationship between  $X$  and  $Y$  (all



**Exhibit 16.6**  
**Least-Squares Regression Line Fitted to Sample Data**

the variation in  $Y$  is explained by the variation in  $X$ ), then  $R^2$  equals 1. At the other extreme, if there is no relationship between  $X$  and  $Y$ , then none of the variation in  $Y$  is explained by the variation in  $X$ , and  $R^2$  equals 0.

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

where  
 Explained variation = Total variation – Unexplained variation

The coefficient of determination for the Stop 'N Go data example is computed as follows. [See Exhibit 16.5 for calculation of  $(Y - \hat{Y})^2$  and  $(Y - \bar{Y})^2$ .]

$$\begin{aligned} R^2 &= \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}} \\ &= 1 - \frac{\text{Unexplained variation}}{\text{Total variation}} \\ &= 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= 1 - \frac{171,605}{871,860} = .803 \end{aligned}$$

Of the variation in  $Y$  (annual sales), 80 percent is explained by the variation in  $X$  (average daily vehicular traffic). There is a very strong linear relationship between  $X$  and  $Y$ .

**Statistical Significance of Regression Results** In computing  $R^2$ , the total variation in  $Y$  was partitioned into two component sums of squares:

$$\text{Total variation} = \text{Explained variation} + \text{Unexplained variation}$$


The total variation is a measure of variation of the observed  $Y$  values around their mean  $\bar{Y}$ . It measures the variation of the  $Y$  values without any consideration of the  $X$  values.

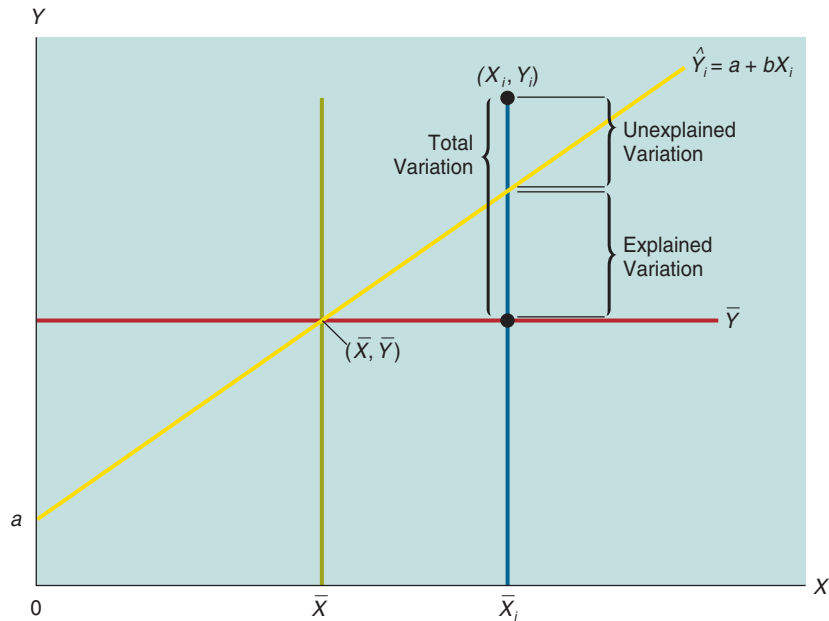
Total variation, known as the *total sum of squares* (SST), is given by

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

The explained variation, or the **sum of squares due to regression** (SSR), is given by

$$\text{SSR} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = a \sum_{i=1}^n Y_i + b \sum_{i=1}^n X_i Y_i - \left( \frac{\sum_{i=1}^n Y_i}{n} \right)^2$$

 **sum of squares due to regression**  
 Variation explained by the regression.



**Exhibit 16.7**  
**Measures of Variation in a Regression**

Exhibit 16.7 depicts the various measures of variation (that is, sum of squares) in a regression. SSR represents the differences between  $Y_i$  (the values of  $Y$  predicted by the estimated regression equation) and  $\bar{Y}$  (the average value of  $Y$ ). In a well-fitting regression equation, the variation explained by regression (SSR) will represent a large portion of the total variation (SST). If  $Y_i = \hat{Y}_i$  at each value of  $X$ , then a perfect fit has been achieved. All the observed values of  $Y$  are then on the computed regression line. Of course, in that case,  $SSR \neq SST$ .

**error sum of squares**  
 Variation not explained by the regression.

The unexplained variation, or **error sum of squares (SSE)**, is obtained from

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n Y_i^2 - a \sum_{i=1}^n Y_i - b \sum_{i=1}^n X_i Y_i$$

In studying the relationship between vehicular traffic and sales, the coefficient of determination may be used to measure the percent of the total variation.



In Exhibit 16.7, note that SSE represents the residual differences (error) between the observed and predicted  $Y$  values. Therefore, the unexplained variation is a measure of scatter around the regression line. If the fit were perfect, there would be no scatter around the regression line and SSE would be zero.

**Hypotheses Concerning Overall Regression** Here we, as the researchers, are interested in hypotheses regarding the computed  $R^2$  value for the problem. Is the amount of variance explained in the result (by our model) significantly greater than we would expect due to chance? Or, as with the various statistical tests discussed in Chapter 15, to what extent can we rule out sampling error as an explanation of the results? Analysis of variance (an  $F$  test) is used to test the significance of the results.

An analysis of variance table is set up as shown in Exhibit 16.8. The computer output for our example appears in Exhibit 16.9. The breakdowns of the total sum of squares and associated degrees of freedom are displayed in the form of an analysis of variance (ANOVA) table. We use the information in this table to test the significance of the linear relationship between  $Y$  and  $X$ . As noted previously, an  $F$  test will be used for this purpose. Our hypotheses are as follows:

- Null hypothesis  $H_0$ : There is no linear relationship between  $X$  (average daily vehicular traffic) and  $Y$  (annual sales).
- Alternative hypothesis  $H_a$ : There is a linear relationship between  $X$  and  $Y$ .

As in other statistical tests, we must choose  $\alpha$ . This is the likelihood that the observed result occurred by chance, or the probability of incorrectly rejecting the null hypothesis. In this case, we decide on a standard level of significance:  $\alpha = .05$ . In other words, if the calculated value of  $F$  exceeds the tabular value, we are willing to accept a 5 percent

EXHIBIT 16.8		Analysis of Variance			
Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F Statistic	
Regression (explained)	1	SSR	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$	
Residual (unexplained)	$n - 2$	SSE	$MSE = \frac{SSE}{n-2}$		
Total	$n - 1$	SST			

EXHIBIT 16.9		Regression Analysis Output				
STAT. MULTIPLE REGRESS.	Regression Summary for Dependent Variable: Y $R = .89619973$ $R^2 = .80317395$ Adjusted $R^2 = .79223917$ $F(1,18) = 73.451$ $p < .00000$ Std. Error of estimate: 97.640					
$N = 20$	BETA	St. Err. of BETA	B	St. Err. of B	$t(18)$	$p$ -level
Intercpt			240.8566	73.38347	3.282164	.004141
X	.896200	.104570	14.7168	1.71717	8.570374	.000000

chance of incorrectly rejecting the null hypothesis. The value of  $F$ , or the  $F$  ratio, is computed as follows (see Exhibit 16.9):

$$F = \frac{MSR}{MSE} = \frac{700,255}{9,534} = 73.5$$

We will reject the null hypothesis if the calculated  $F$  statistic is greater than or equal to the table, or critical,  $F$  value. The numerator and denominator degrees of freedom for this  $F$  ratio are 1 and 18, respectively. As noted earlier, it was decided that an alpha level of .05 ( $\alpha = .05$ ) should be used.

The table, or critical, value of  $F$  with 1 (numerator) and 18 (denominator) degrees of freedom at  $\alpha = .05$  is 4.49 (see Table 5 in Appendix 2). Because the calculated value of  $F$  is greater than the critical value, we reject the null hypothesis and conclude that there is a significant linear relationship between the average daily vehicular traffic ( $X$ ) and annual sales ( $Y$ ). This result is consistent with the high coefficient of determination  $R^2$  discussed earlier.

**Hypotheses about the Regression Coefficient  $b$**  Finally, we may be interested in making hypotheses about  $b$ , the regression coefficient. As you may recall,  $b$  is the estimate of the effect of a one-unit change in  $X$  on  $Y$ . The hypotheses are as follows:

- Null hypothesis  $H_0$ :  $b = 0$ .
- Alternative hypothesis  $H_a$ :  $b \neq 0$ .

The appropriate test is a  $t$  test, and, as you can see from the last line of Exhibit 16.9, the computer program calculates the  $t$  value (8.57) and the  $p$  value (probability of incorrectly rejecting the null hypothesis of .0000). See Chapter 15 for a more detailed discussion of  $p$  values. Given the  $\alpha$  criterion of .05, we would reject the null hypothesis in this case.

## Correlation Analysis

### Correlation for Metric Data: Pearson's Product-Moment Correlation

Correlation is the degree to which changes in one variable (the dependent variable) are associated with changes in another. When the relationship is between two variables, the analysis is called simple, or bivariate, **correlation analysis**. With metric data, **Pearson's product-moment correlation** may be used.

In our example of bivariate regression, we used the coefficient of determination  $R^2$  as a measure of the strength of the linear relationship between  $X$  and  $Y$ . Another descriptive measure, called the *coefficient of correlation*  $R$ , describes the degree of association between  $X$  and  $Y$ . It is the square root of the coefficient of determination with the appropriate sign (+ or -):

$$R = \pm \sqrt{R^2}$$

The value of  $R$  can range from  $-1$  (perfect negative correlation) to  $+1$  (perfect positive correlation). The closer  $R$  is to  $\pm 1$ , the stronger the degree of association between  $X$  and  $Y$ . If  $R$  is equal to zero, then there is no association between  $X$  and  $Y$ .

**correlation analysis**  
Analysis of the degree to which changes in one variable are associated with changes in another.

**Pearson's product-moment correlation**  
Correlation analysis technique for use with metric data.

# PRACTICING MARKETING RESEARCH



## A Novel Application: Using the Pearson Product-Moment Correlation to Study Sayings in the Gospel of Thomas

Since the 1970s, biblical scholars have been vexed as to whether the Gospel of Thomas used any of the synoptic gospels as a source for its 100 sayings attributed to Jesus. Using the laborious scholarly method respectfully called careful learned consideration, scholars sought to

correlation, Davies created two statistical tables, based on the 100 with no doubtful parallels.

The first table shows the correlations between the sayings found in Thomas that are also found in the gospels to their order in those texts. Davies explains that in both tables, the Degree of Correlation ( $R$ ) shows the closeness of linear relationship between the two named variables, and that the Reliability of Correlation ( $p$ ) shows how much  $R$  was likely due to chance. That means that the lower  $p$  is, the lower it is likely that the correlation  $R$  observed was due to chance.  $N$  stands for the number of pairs of sayings that Davies analyzed.

### Matthew parallels in Thomas to Luke parallels in Thomas

$R = .528$   $N = 62$   
 $p = .0001$

### Matthew parallels in Thomas to Mark parallels in Thomas

$R = .876$   $N = 29$   
 $p = .0001$

### Luke parallels in Thomas to Mark parallels in Thomas

$R = .709$   $N = 25$   
 $p = .0001$

### Matthew parallels in Thomas to Luke parallels in Thomas

$R = .460$   $N = 37$   
 $p = .004$

compare the order of the quoted sayings in Thomas against that found in the gospels to see if there is a direct relationship. But this is “fundamentally a question for technical statistical analysis,” decided Stevan L. Davies, professor of Religious Studies at College Misericordia in Dallas, Pennsylvania, and translator of *The Gospel of Thomas*.

Davies worked with two lists. One had 110 sayings but with some doubtful parallels; the other had 100, with no doubtful parallels. In all he would compare the sayings from Thomas with about 226 other ordered sayings in the gospels. Using the Pearson product-moment

According to Davies, the correlations displayed in the first table are not likely to be due to chance. The sayings used in this first table came from a random list, namely, only those found in Thomas, and are important to the study mainly because they confirm the applicability of the correlation method used.

In the second table, Davies shows how the method demonstrates what the biblical scholars had presumed for 30 years, that there is no statistically significant correlation in the order of the paralleled sayings in Thomas with the order of sayings in the other gospels or even to any source-subset of the gospels.<sup>2</sup>

### Mark parallels in Thomas's order to Mark's order

$R = .213$   
 $N = 30$   
 $p = .258$

### All Matthew parallels in Thomas's order to Matthews' order

$R = .088$   
 $N = 82$   
 $p = .431$

### All Luke parallels in Thomas's order to Luke's order

$R = .141$   
 $N = 77$   
 $p = .221$

### Matthew parallels in Thomas's order to Matthews' order

$R = .062$   
 $N = 37$   
 $p = .714$

### Luke parallels in Thomas's order to Luke's order

$R = .074$   
 $N = 37$   
 $p = .663$

### Matthew parallels in Thomas's order to Matthews' order

$R = .084$   
 $N = 25$   
 $p = .689$

### Luke parallels in Thomas's order to Luke's order

$R = .170$   
 $N = 25$   
 $p = .415$



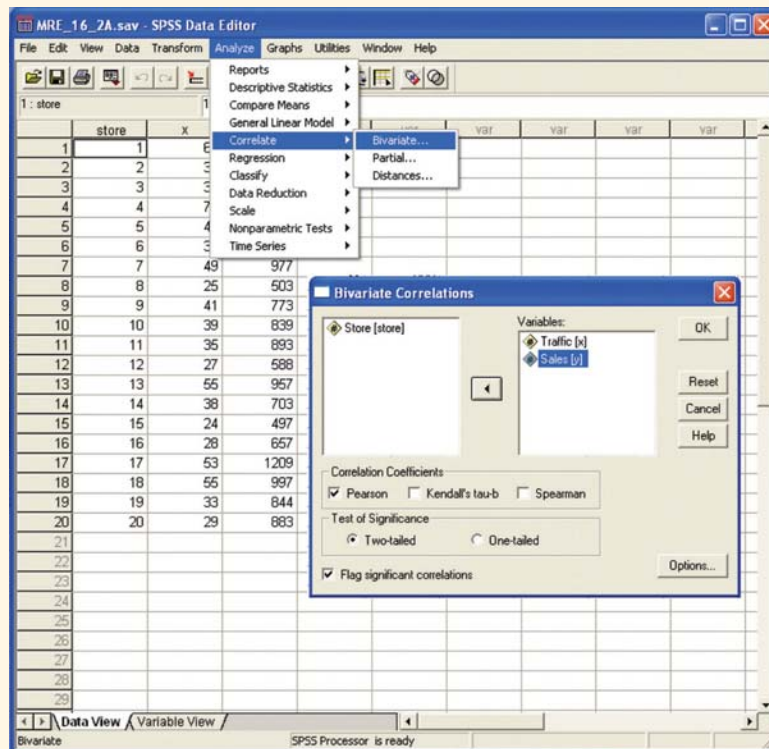
If we had not been interested in estimating the regression function, we could have computed  $R$  directly from the data for the convenience store example, using this formula:

$$\begin{aligned}
 R &= \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} \\
 &= \frac{20(734,083) - (816)(16,826)}{\sqrt{[20(36,526) - (816)^2][20(15,027,574) - (16,826)^2]}} \\
 &= .896
 \end{aligned}$$

In this case, the value of  $R$  indicates a positive correlation between the average daily vehicular traffic and annual sales. In other words, successively higher levels of sales are associated with successively higher levels of traffic.

## SPSS JUMP START FOR CORRELATION

Steps that you need to go through to do the correlation problem shown in the book are provided below along with the output produced. Use the data set **Correx**, which you can download from the Web site for the text.



## Steps in SPSS

1. Select *Analyze* → *Correlate* → *Bivariate*.
2. Move **x** to Variables.
3. Move **y** to Variables.
4. Click OK.

# SPSS Output for Correlation

## Correlations

### Correlations

		Traffic	Sales
Traffic	Pearson Correlation	1	.896**
	Sig. (2-tailed)	.	.000
	N	20	20
Sales	Pearson Correlation	.896**	1
	Sig. (2-tailed)	.000	.
	N	20	20

\*\* . Correlation is significant at the 0.01 level

## PRACTICING MARKETING RESEARCH



### ***When Bivariate and Multi-variate Correlation Combined Give the Best Picture***

Social researchers Douglas Kirby, Karin Coyle, and Jeffrey B. Gould wanted to assess the relationships between conditions of poverty and birthrates among young teenagers in California. They collected data from 1,811 zip codes in which any teenage births had been recorded between 1991 and 1996. They excluded all zip code areas that did not have at least 200 young women aged 15–17 (which they called “young teenage birthrates”) to get a sample of 1,192 zip codes.

Their dependent variable was the mean of the yearly birthrates for women in this group. Their independent variables included 19 demographic features, which they culled from a list of 177 social indicators. Of these 19 independent measures, 3 dealt with ethnicity and 16 represented other factors such as education, employment, marital status, income level, and housing status.

Using these data, the researchers calculated the simple bivariate correlation and regression coefficients between young teenage birthrate and the 19 social measures, one at a time. Their bivariate analysis results showed that the number of families living in conditions of poverty in a

given zip code was “highly related” to the birthrate among teenagers 15–17. The bivariate correlations, they concluded, “show that a single variable, the proportion of households living below the poverty line, is highly related to the young-teenage birthrate.” Bivariate analysis also demonstrated that median household income and the number of homes receiving public assistance are also highly related and that three of the four poverty measures have the largest regression coefficients.

But the researchers wanted to look at a bigger picture of relationships and control for the “correlates” of family poverty level. So they shifted to multivariate correlation to make connections among multiple manifestations of poverty, low educational levels and employment status, and high levels of employment. They found that these factors also have a “large impact” on teenage birthrates. Multivariate correlation showed that the number of families living at or below poverty levels “remained by far the most important predictor” of teenage birthrate.<sup>3</sup>

Similarly, researcher and author Clayton E. Cramer found that the sequential application of bivariate and then multivariate correlation produced the best results in his study of the effectiveness of the Brady Handgun Violence Prevention Act of 1993. Bivariate analysis is easy to perform, Cramer says, and works well for certain types of research problems, such as comparing brands of gun ammunition or suggesting that factor A did not cause factor B or that factor A affected factor B.

But when you tackle “hard social problems” such as those associated with crime and gun control, using only two variables is insufficient for figuring out true causality. “Unlike bivariate correlation analysis, multivariate correlation analysis can help identify some truly subtle relationships—where a 3 percent increase in A may cause a 1 percent increase in B.” Multivariate analysis is “a devilishly complex technique,” and scientists using it can make legitimate mistakes hard to detect except by other scientists, Cramer says, but application of it produced strong data that the Brady Law had no effect on homicide rates.<sup>4</sup>

## SUMMARY

The techniques used to analyze the relationship between variables taken two at a time are called bivariate analyses. Bivariate regression analysis allows a single dependent variable to be predicted from knowledge about a single independent variable. One way to examine the underlying relationship between a dependent and an independent variable is to plot them on a scatter diagram. If the relationship appears to be linear, then linear regression analysis may be used. If it is curvilinear, then curve-fitting techniques should be applied. The general equation for a straight line fitted to two variables is given by

$$Y = a + bX$$

where  $Y$  = dependent variable  
 $X$  = independent variable  
 $a$  =  $Y$  intercept  
 $b$  = amount  $Y$  increases with each unit increase in  $X$

Both  $a$  and  $b$  are unknown and must be estimated. This process is known as simple linear regression analysis. Bivariate least-squares regression analysis is a mathematical technique for fitting a line to measurements of the two variables  $X$  and  $Y$ . The line is fitted so that the algebraic sum of deviations of the actual observations from the line is zero and the sum of the squared deviations is less than it would be for any other line that might be fitted to the data.

The estimated regression function describes the nature of the relationship between  $X$  and  $Y$ . In addition, researchers want to know the strength of the relationship between the variables. This is measured by the coefficient of determination, denoted by  $R^2$ . The coefficient of determination measures the percent of the total variation in  $Y$  that is “explained” by the variation in  $X$ . The  $R^2$  statistic ranges from 0 to 1. An analysis of variance (ANOVA) approach also can be used for regression analysis. The total variation is known as the total sum of squares (SST). The explained variation, or the sum of squares due to regression (SSR), represents the variability explained by the regression. The unexplained variation is called the error sum of squares (SSE).

Correlation analysis is the measurement of the degree to which changes in one variable are associated with changes in another. Correlation analysis will tell the researcher whether the variables are positively correlated, negatively correlated, or independent.

**bivariate techniques** Statistical methods of analyzing the relationship between two variables.

**independent variable** Variable believed to affect the value of the dependent variable.

**dependent variable** Variable expected to be explained or caused by the independent variable.

**bivariate regression analysis** Analysis of the strength of the linear relationship between two variables when one is considered the independent variable and the other the dependent variable.

**coefficient of determination** Percentage of the total variation in the dependent variable explained by the independent variable.

**sum of squares due to regression** Variation explained by the regression.

**error sum of squares** Variation not explained by the regression.

**correlation analysis** Analysis of the degree to which changes in one variable are associated with changes in another.

**Pearson’s product-moment correlation** Correlation analysis technique for use with metric data.

## KEY TERMS & DEFINITIONS

1. Give an example of a marketing problem for which use of each of the three procedures listed in question 1 would be appropriate.
2. A sales manager of a life insurance firm administered a standard multiple-item job satisfaction scale to all the members of the firm’s salesforce. The manager then correlated (Pearson’s product-moment correlation) job satisfaction score with years of school completed for each salesperson. The resulting correlation was .11. On the basis of this analysis, the sales manager concluded: “A salesperson’s level of education has little to do with his or her job satisfaction.” Would you agree or disagree with this conclusion? Explain the basis for your position.
3. What purpose does a scatter diagram serve?
4. Explain the meaning of the coefficient of determination. What does this coefficient tell the researcher about the nature of the relationship between the dependent and independent variables?

## QUESTIONS FOR REVIEW & CRITICAL THINKING

5. It has been observed in the past that when an AFC team wins the Super Bowl, the stock market rises in the first quarter of the year in almost every case. When an NFC team wins the Super Bowl, the stock market falls in the first quarter in most cases. Does this mean that the direction of movement of the stock market is caused by which conference wins the Super Bowl? What does this example illustrate?
6. The following table gives data collected by a convenience store chain for 20 of its stores.

Column 1: ID number for each store

Column 2: Annual sales for the store for the previous year in thousands of dollars

Column 3: Average number of vehicles that pass the store each day, based on actual traffic counts for one month

Column 4: Total population that lives within a 2-mile radius of the store, based on 1990 census data

Column 5: Median family income for households within a 2-mile radius of the store, based on 2000 census data

Store ID No.	Annual Sales (thousands of dollars)	Average Daily Traffic	Population in 2-Mile Radius	Average Income in Area
1	\$1,121	61,655	17,880	\$28,991
2	\$ 766	35,236	13,742	\$14,731
3	\$ 595	35,403	19,741	\$ 8,114
4	\$ 899	52,832	23,246	\$15,324
5	\$ 915	40,809	24,485	\$11,438
6	\$ 782	40,820	20,410	\$11,730
7	\$ 833	49,147	28,997	\$10,589
8	\$ 571	24,953	9,981	\$10,706
9	\$ 692	40,828	8,982	\$23,591
10	\$1,005	39,195	18,814	\$15,703
11	\$ 589	34,574	16,941	\$ 9,015
12	\$ 671	26,639	13,319	\$10,065
13	\$ 903	55,083	21,482	\$17,365
14	\$ 703	37,892	26,524	\$ 7,532
15	\$ 556	24,019	14,412	\$ 6,950
16	\$ 657	27,791	13,896	\$ 9,855
17	\$1,209	53,438	22,444	\$21,589
18	\$ 997	54,835	18,096	\$22,659
19	\$ 844	32,916	16,458	\$12,660
20	\$ 883	29,139	16,609	\$11,618

Answer the following:

- a. Which of the other three variables is the best predictor of sales? Compute correlation coefficients to answer the question.
  - b. Do the following regressions:
    1. Sales as a function of average daily traffic
    2. Sales as a function of population in a 2-mile radius
  - c. Interpret the results of the two regressions.
7. Interpret the following:
- a.  $Y = .11 + .009X$ , where  $Y$  is the likelihood of sending children to college and  $X$  is family income in thousands of dollars. Remember: It is family income in *thousands*.

1. According to our model, how likely is a family with an income of \$100,000 to send their children to college?
  2. What is the likelihood for a family with an income of \$50,000?
  3. What is the likelihood for a family with an income of \$17,500?
  4. Is there some logic to the estimates? Explain.
- b.  $Y = .25 - .0039X$ , where  $Y$  is the likelihood of going to a skateboard park and  $X$  is age.
1. According to our model, how likely is a 10-year-old to go to a skateboard park?
  2. What is the likelihood for a 60-year-old?
  3. What is the likelihood for a 40-year-old?
  4. Is there some logic to the estimates? Explain.
8. The following ANOVA summary data are the result of a regression with sales per year (dependent variable) as a function of promotion expenditures per year (independent variable) for a toy company.

$$F = \frac{MSR}{MSE} = \frac{34,276}{4,721}$$

The degrees of freedom are 1 for the numerator and 19 for the denominator. Is the relationship statistically significant at  $\alpha = .05$ ? Comment.

1. Go to <http://www.grapentine.com/displayrn.asp?Id=17> for an excellent discussion of the role that collinearity plays in regression analysis. A variety of other statistical issues in a marketing context are discussed at this site.
2. Go to <http://www.grapentine.com/displayrn.asp?Id=54> for a very good discussion of interpretation of the correlation coefficient.
3. Go to <http://www.spss.com/regression/> for discussion of regression applications provided by SPSS.

## WORKING THE NET

## Axcis Athletic Shoes

Fred Luttrell is the new product development manager for Axcis Athletic Shoe Company. He recently completed consumer testing of 12 new shoe concepts. As part of this test, a panel of consumers was asked to rate the 12 shoe concepts on two attributes, overall quality and style. A 10-point scale was used with anchors at 10 = best possible and 1 = worst possible.

The panel of 20 consumers met as a group and came up with the ratings as a group. Fred believes that there is a relationship between the style ratings and the overall quality

## REAL-LIFE RESEARCH • 16.1

ratings. He believes that shoes receiving higher ratings on style also will tend to receive higher ratings on overall quality. The ratings results for the 12 shoe concepts are as follows.

Shoe Model	Style Rating	Quality Rating
1	9	8
2	7	7
3	6	8
4	9	9
5	8	7
6	5	5
7	9	7
8	7	9
9	8	6
10	10	9
11	6	5
12	9	10

### Questions

1. Which of the statistical procedures covered in this chapter is appropriate for addressing Fred's theory? Why would you choose that technique over the other?
2. Use the technique that you chose to determine whether Fred's theory is supported by the statistical evidence. State the appropriate null and alternative hypotheses. Is Fred's theory supported by the statistical evidence? Why or why not?

## REAL-LIFE RESEARCH • 16.2

### Find Any Error?

*Bloomberg Personal* is a monthly investment magazine published for upscale U.S. retail investors. Each issue of *Bloomberg Personal* now includes a section entitled "Global Reach," which gives a brief description of activity in the national stock and bond markets of foreign countries. The performance of foreign financial markets can be useful in gauging current economic conditions for consumers in these foreign countries.

In the December 1997 issue, nine countries were featured, including Japan, Australia, Russia, and China.<sup>5</sup> In addition, an inset box included the following information:

Here's how closely some markets have mirrored the S & P 500 (a stock index of 500 leading firms in the United States) since 1992. The closer the figure is to 1.0, the higher the correlation.

Canada	.62
UK	.41
France	.39
Germany	.37
Italy	.22
Singapore	.19
Chile	.15
Brazil	.14
Philippines	.08
Turkey	-.02

Source: "A Mirror to the World," *Bloomberg Personal*, December 1997. Reprinted by permission.

## Questions

1. What other information would be essential to have before you could use these secondary data in a marketing research report?
2. If you had this other information and the precision of measurement was acceptable to you, what is the story these numbers would begin to tell about the linkages between the U.S. stock market and the stock markets of some other nations in the world?



## SPSS EXERCISES FOR CHAPTER 16

*Note:* If you did not complete any of the SPSS exercises in Chapter 14, you will need a corrected database from your professor.

### Exercise #1: Bivariate Regression

Use the *analyze/regression/linear* sequence to invoke bivariate regression analysis. This exercise attempts to explain the variation in the *number of movies the respondent attends in an average month (Q3)*. Hence, **Q3** is the **dependent variable**. Invoke the bivariate regression procedure for the following pairs of variables:

1. Q3 and Q5d (movie theater item—importance of comfortable chairs)
2. Q3 and Q5e (movie theater item—auditorium type seating)
3. Q3 and Q7a (movie theater information source—newspaper)
4. Q3 and Q7b (movie theater information source—Internet)
5. Q3 and Q7c (movie theater information source—phone in for information)
6. Q3 and Q9 (self-perception of how physically active)
7. Q3 and Q10 (self-perception of how socially active)

Summarize the results of the bivariate regression analysis, by filling in tables similar to the ones below.

Model	Regression coefficient	t	Sig.
Constant			
Q5d			
Q5e, etc.			

Variables	Model R <sup>2</sup>	Model F-value	Sig.
Q5d			
Q5e, etc.			





1. At the 95 percent level of confidence, which of the regression models (list the pairs of variables) are significant (list the dependent variables)?  
\_\_\_\_\_
2. **Interpretation of the regression coefficients:** Use the following table to summarize the regression coefficient,  $b$ , in each of the 7 regression models.

Model	Regression Coefficient $b$	$t$	Sig. of $b$	Interpretation of the Regression Coefficient $b$
Example Q3 & Q5b	.244	4.147	.000	A one-unit increase in Q5b is associated with a .244 increase in monthly movie attendance

3. Using the regression results, compute  $Y(Q3)$  if  $Q5d = 4$ . \_\_\_\_\_
4. Using the regression results, compute  $Y(Q3)$  if  $Q7c = 2$ . \_\_\_\_\_
5. Using the regression results, compute  $Y(Q3)$  if  $Q9 = 3$ . \_\_\_\_\_
6. Which of the 7 models in the bivariate regression analysis explained the most variation in Q3 (*hint:  $R^2$* )? \_\_\_\_\_
7. In which of the 7 models does the independent variable's regression coefficient cause the largest change in Q3 for a one unit change in the independent variable? \_\_\_\_\_

## Exercise #2: Pearson's Product-Moment Correlation

Use the *analyze/correlate/bivariate* sequence to invoke bivariate correlation analysis. This exercise utilizes the metric correlation technique (Pearson's), which requires that both variables in the bivariate analysis be of at least interval measurement scale. The objective of this exercise is to examine the association between various pairs of variables.

Invoke the bivariate correlation procedure utilizing the Pearson coefficient to evaluate the association between the following pairs of variables:

- a. Q3 and Q8a (Purchase Option for movie tickets—Internet)
- b. Q9 (self-perception of how physically active) and Q10 (self-perception of how socially active)
- c. Q8a (Purchase Option for movie tickets—Internet) and Q7b (importance of the Internet as a source of information about movies at movie theaters)
- d. Q5b (Movie Theater Item—importance of soft drinks and food items) and Q9 (self-perception of how physically active)
- e. Q5h (Movie Theater Item—number of screens at a movie theater) and Q10 (self-perception of how socially active)

With the results of the bivariate correlation using the Pearson coefficient, fill in a table similar to the one below.

Variables	Pearson Coefficient (include +/-)	Probability of an insignificant correlation in the population (based on the sample results)	Interpretation of the results

Questions to Answer: (Assume a significant relationship requires at least a **95 percent** level of confidence.)

1. Of the five correlations computed, which pair of variables had the strongest association? \_\_\_\_\_
2. Of the three correlations computed, which pair of variables had the weakest association? \_\_\_\_\_
3. Do people who perceive themselves as more physically active have a greater or lesser need for food and drink at a movie theater? \_\_\_\_\_
4. Are people who use the Internet to purchase movie tickets more or less likely to use the Internet to get information about movies at movie theaters? \_\_\_\_\_

