CHAPTER 18

Measures of Association

Learning Objectives

After reading this chapter, you should understand . . .

1. How correlation analysis may be applied to study relationships between two or more variables.
2. The uses, requirements, and interpretation of the product moment correlation coefficient.
3. How predictions are made with regression analysis using the method of least squares to minimize errors in drawing a line of best fit.
4. How to test regression models for linearity and whether the equation is effective in fitting the data.
5. The nonparametric measures of association and the alternatives they offer when key assumptions and requirements for parametric techniques cannot be met.

Bringing Research to Life

Myra arrived for an analysis meeting with Jason and found a round, bald little man sitting at Jason's desk, studying the screen of a laptop computer, stroking his gray beard and smiling broadly.

"Myra," said Jason, "meet my Uncle Jack."

"Really? At last I meet the famous—or should I say infamous—uncle? I feared you were imaginary."

"Reports of my being imaginary are greatly exaggerated, if I may paraphrase Mark Twain."

It seemed to her he should have been taller. The ex-biggest painting contractor on Long Island ought to be more ample, she felt. Maybe he had shrunk upon retiring to Florida. Uncle Jack was the kind of unassuming diminutive old fellow who often proved to have money, or power, or both, though no one would have guessed.

Uncle Jack began to rub his laptop computer and grin even more broadly. "I wanted Jason to see that I am taking *good care* of this MindWriter." Myra thought it was a great "box," but what made this one so special?

"This little computer," said Uncle Jack, "has made me the political kingpin of the Boca Beach Condominium Association, Phases One *and* Two. Not to mention all the widows who want to meet with me. Which I had better not mention, as I have only been widowed for a short while and don't want this getting back to my boys on Long Island. I gave the painting business to my three boys and—hey, I'm 75 and deserve some fun—I moved to Boca Raton. For three months I played golf in the morning and sat by the pool and played cards in the afternoon. For three months, seven days, I did this. Do I have to tell you? I was going crazy.

"Then my next-door neighbor Marty 'headed south,' and his wife gave me his MindWriter."

Myra tried to imagine what was south of Florida and why Marty would go there without his wife. Uncle Jack saw the confusion and explained, "She 'planted him,' is what I mean. Do you follow me? He died, he met his 'necessary end,' if I may quote from Julius Caesar. And what was a widow going to do with this computer?

"So now I had a laptop, and what was I going to do with it? Jason came through Boca Raton and copied me a program from his computer, but it was a statistical program, free from the Internet. That's all he had that he could give me without breaking the copyright law, a statistical program from the Internet. And me the guy who had to take statistics three times at Brooklyn College in 1948, to get a lousy C+.

"Something developed pretty soon. We had a dirty guy, Sandy Plover, who had come down from being a big electrical contractor in Jersey and got himself into condo politics. Sandy had not become successful up North without developing a sense of who holds power, what their weaknesses are, and how the politicians agitate the electorate. Being a natural-born troublemaker, he waited for his chance to agitate. Well, the sheriff had launched a statistics-reporting program, and the statistical results had started to come in. Now, it happens there is a creek that divides our condominium into two halves of roughly equal population, Oceanside and Gladeside, the former being slightly tonier. And it further happens that the incidence of arrests resulting from police calls to Oceanside, where he lives, is higher than in Gladeside."

Uncle Jack wrote the following:

H_1: Gladeside residents get special treatment when it comes to solving crimes and thus live in a safer environment due to their higher incomes and greater political power.

H_0: Gladeside and Oceanside receive the same attention from the police.

In Gladeside:		
Police calls without arrest	46	
Police calls resulting in arrests	4	50
In Oceanside:		
Police calls without arrest	40	
Police calls resulting in arrests	10	50

"I doubt that Sandy would have paid attention, except that in both sides of the condo association the total number of police calls happened to be 50, which made it easy for him to see that in Oceanside the rate of arrests was twice that in Gladeside."

"Actually," said Myra, "I'm surprised there would be any police calls in such a tony condo community."

"We are old," said Uncle Jack, "but not dead.

"In any case, Sandy's finely honed political instincts told him he was going to go nowhere by trying to turn the condo against the sheriff. It would be much, much better, he saw, to try and turn voters of Oceanside against those in Gladeside. And so he decided to complain about the disparate impact of arrests, though being self-educated, he had never heard that expression, 'disparate.' All he knew is that here was an opportunity to make trouble and thereby make a name, because in Oceanside were mostly folks moved down from Brooklyn, and in Gladeside folks from the Bronx, and there was an undercurrent, if you know what I mean."

"But the ethics . . ."

". . . meant nothing to Sandy. He told me, 'I think I am gonna kick some butt and make a name for myself down here.'

"Right away, you can see his strategy. Sandy thought he would make more mileage by whipping our side of the association against the other. This had something to do with his wanting to run in a jurisdiction where elections were not done at large but on a single-district basis, and something to do with his rabble-rousing instincts, which had always been impeccable."

Jason interrupted with a minilecture on the politics of statistics. "The trouble with the police calls as an issue is that sheriffs' offices nowadays are well staffed with statistically educated analysts who know very well how to rebut oddball claims."

"Personally, I miss the old days, when if you worked for the cops, you busted bad guys' heads and never mind statistics. But I understand nowadays you don't mess with crime statistics without one of the pencil-heads rebutting you. So I punched the numbers into this MindWriter here to double-check the stats. I did the obvious first, just what I supposed a police analyst would have done. I ran a cross-tabulation and a chi-square test of the hypothesis that the arrests in Oceanside were disproportionate to those in Gladeside."

"But," said Myra, "you only have to look at the numbers to see they are."

"Yes, but what you have there is the 'eyeball' fallacy, as my dear old professor called it almost 50 years ago. As I explained to Sandy, 100 police encounters resulting in a few arrests is nothing, nada, not a large enough sample to trust a quick peek and a leap to a conclusion. You run it through the computer, and, sure enough, although the ratios seem to be out of whack, they are not statistically significant. You cannot support disparate impact. No way."

Jason said, "Granting that 10 arrests per 50 is bigger than 4 per 50, Uncle Jack saw that a statistician would say that it is not significantly bigger, would say that it is not disproportionate enough to convince a *scientist* that the police were acting differently in the two sides of town. A statistician would say, wait and see, let the story unfold, collect a bigger sample."

"How did Sandy accept your explanation, Jack?"

"Like the mad dog he was. To quote my favorite poet, 'There are chords in the hearts of the most reckless which cannot be touched without emotion.'"

"Emerson?" asked Myra.

"Poe," said Uncle Jack. "Underrated as a poet, overrated as a mystery writer. Well, as for Sandy, he was ready to shoot the messenger, very much bothered, at first, that I would not support his political strategy, and he grumbled that I did not understand such things, that you had to do bad to do good, which I am not sure I agree with, and that I was maybe too much overeducated to ever understand the need just to get on with investigating issues. Me, overeducated, an English major!

"But I was not interested in right and wrong, I said, and he could be pretty sure the sheriff would come roaring back with a statistical analysis to throw cold water over Sandy."

"Did you bring him around?"

"That jerk, come around? Never. He ran to the papers, they splashed his numbers all over the third page of the local section on Monday, and Tuesday the sheriff came back with his experts and made Sandy look like a fool . . . except that it was a slow news day on Tuesday and the paper gave the story plenty of ink, on page one, if you can believe it. So Sandy was washed up, and I am now the resident genius. What I do is look at the opponents' polling results and deny their validity for the newspapers and TV. If the opposing party is ahead by a few poll points, I scoff at the thinness of the margin. If their lead is wide, I belittle the size of the sample and intimate that any statistician would see through them."

Jason provided the scholarly footnote. "Uncle Jack is colorful, amusing, and good-natured in debunking his opponents' polls, and the newspaper writers, who understand less statistics than anyone, have never challenged him to substantiate his claims. What he learned from me is that statistics is so complicated, and scares so many people, that you can claim or deny anything. And he is usually right to debunk the polls, since for a preelection political poll to be taken seriously there has to be a large enough sample to produce significant results. And there has to be a big enough spread between winners and losers to protect against a last-minute shift in voter sentiment. In the small, closely contested voting precincts of condominium politics, hardly any poll can meet two such stringent criteria."

"Right, Jason. I sit in the clubhouse and people come over and want to know what I think about the Middle East, campaign reform, everything. When you have a computer, to paraphrase Tevya, 'they think you really know.'"

Introduction

You might want to review our discussion of relational hypotheses in Chapter 2.

In the previous chapter, we emphasized testing hypotheses of *difference*. However, management questions in business frequently revolve around the study of relationships between two or more variables. Then, a *relational hypothesis* is necessary. In the research question "Are U.S. kitchen appliances perceived by American consumers to be of better quality than foreign kitchen appliances?" the nature of the relationship between the two variables ("country of origin" and "perceived quality") is not specified. The implication, nonetheless, is that one variable is responsible for the other. A correct relational hypothesis for this question would state that the variables occur together in some specified manner without implying that one causes the other.

Various objectives are served with correlation analysis. The strength, direction, shape, and other features of the relationship may be discovered. Or tactical and strategic questions may be answered by predicting the values of one variable from those of another. Let's look at some typical management questions:

- In the mail-order business, excessive catalog costs quickly squeeze margins. Many mailings fail to reach receptive or active buyers. What is the relationship between various categories of mailings that delete inactive customers and the improvement in profit margins?
- Medium-sized companies often have difficulty attracting the cream of the MBA crop, and when they are successful, they have trouble retaining them. What is the relationship between the ranking of candidates based on executive interviews and the ranking obtained from testing and assessment?
- Retained cash flow, undistributed profits plus depreciation, is a critical source of funding for equipment investment. During a period of decline, capital spending suffers. What is the relationship between retained cash flow and equipment investment over the last year? Between cash flow and dividend growth?
- Aggressive U.S. high-tech companies have invested heavily in the European chip market, and their sales have grown 20 percent over the three largest European firms. Can we predict next year's sales based on present investment?

All these questions may be evaluated by means of measures of association. And all call for different techniques based on the level at which the variables were measured or the intent of the question. The first three use nominal, ordinal, and interval data, respectively. The last one is answered through simple bivariate regression.

With correlation, one calculates an index to measure the nature of the relationship between variables. With regression, an equation is developed to predict the values of a dependent variable. Both are affected by the assumptions of measurement level and the distributions that underlie the data.

Exhibit 18–1 lists some common measures and their uses. The chapter follows the progression of the exhibit, first covering bivariate linear correlation, then simple regression, and concluding with nonparametric measures of association. Exploration of data through visual inspection and diagnostic evaluation of assumptions continues to be emphasized.

Bivariate Correlation Analysis

Bivariate correlation analysis differs from nonparametric measures of association and regression analysis in two important ways. First, parametric correlation requires two continuous variables measured on an interval or ratio scale. Second, the coefficient does not distinguish between independent and dependent variables. It treats the variables symmetrically since the coefficient r_{xy} has the same interpretation as r_{yx}.

Pearson's Product Moment Coefficient *r*

The **Pearson** (product moment) **correlation coefficient** varies over a range of +1 through 0 to –1. The designation r symbolizes the coefficient's estimate of linear association based on sampling data. The coefficient ρ represents the population correlation.

Correlation coefficients reveal the magnitude and direction of relationships. The *magnitude* is the degree to which variables move in unison or opposition. The size of a correlation of +.40 is the same as one of –.40. The sign says nothing about size. The degree of correlation is modest. The coefficient's sign signifies the *direction* of the relationship. Direction tells us whether large values on one variable are associated with large values on the other (and small values with small values). When the values correspond in this way, the two variables have a positive relationship: As one increases, the other also increases. Family income, for example, is positively related to household

EXHIBIT 18–1 Commonly Used Measures of Association

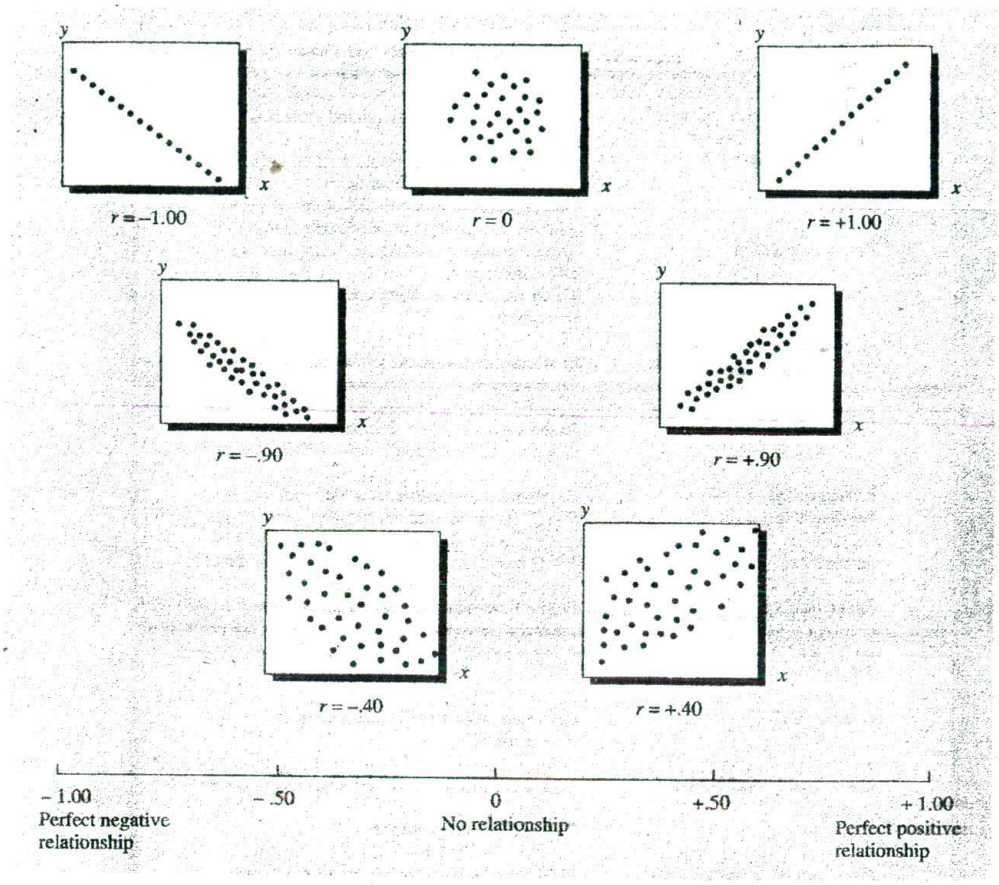
Measurement	Coefficient	Comments and Uses
Interval and Ratio	**Pearson (product moment) correlation coefficient**	For continuous linearly related variables.
	Correlation ratio (eta)	For nonlinear data or relating a main effect to a continuous dependent variable.
	Biserial	One continuous and one dichotomous variable with an underlying normal distribution.
	Partial correlation	Three variables; relating two with the third's effect taken out.
	Multiple correlation	Three variables; relating one variable with two others.
	Bivariate linear regression	Predicting one variable from another's scores.
Ordinal	**Gamma**	Based on concordant-discordant pairs: $(P - Q)$; proportional reduction in error (PRE) interpretation.
	Kendall's tau *b*	$P - Q$ based; adjustment for tied ranks.
	Kendall's tau *c*	$P - Q$ based; adjustment for table dimensions.
	Somers's *d*	$P - Q$ based; asymmetrical extension of gamma.
	Spearman's rho	Product moment correlation for ranked data.
Nominal	**Phi**	Chi-square (CS) based for 2×2 tables.
	Cramer's *V*	CS based; adjustment when one table dimension > 2.
	Contingency coefficient *C*	CS based; flexible data and distribution assumptions.
	Lambda	PRE-based interpretation.
	Goodman & Kruskal's tau	PRE-based with table marginals emphasis.
	Uncertainty coefficient	Useful for multidimensional tables.
	Kappa	Agreement measure.

food expenditures. As income increases, food expenditures increase. Other variables are inversely related. Large values on the first variable are associated with small values on the second (and vice versa). The prices of products and services are inversely related to their scarcity. In general, as products decrease in available quantity, their prices rise. The absence of a relationship is expressed by a coefficient of approximately zero.

Scatterplots for Exploring Relationships

Scatterplots are essential for understanding the relationships between variables. They provide a means for visual inspection of data that a list of values for two variables cannot. Both the direction and the shape of a relationship are conveyed in a plot. With a little practice, the magnitude of the relationship can be seen.

EXHIBIT 18–2 Scatterplots of Correlations between Two Variables



MANAGEMENT Tip

Exhibit 18–2 contains a series of scatterplots that depict some relationships across the range r. The three plots on the left side of the figure have their points sloping from the upper left to the lower right of each x–y plot.[1] They represent different magnitudes of negative relationships. On the right side of the figure, three plots have opposite patterns and show positive relationships.

When stronger relationships are apparent (for example, the ±.90 correlations), the points cluster close to an imaginary straight line passing through the data. The weaker relationships (±.40) depict a more diffuse data cloud with points spread farther from the line.

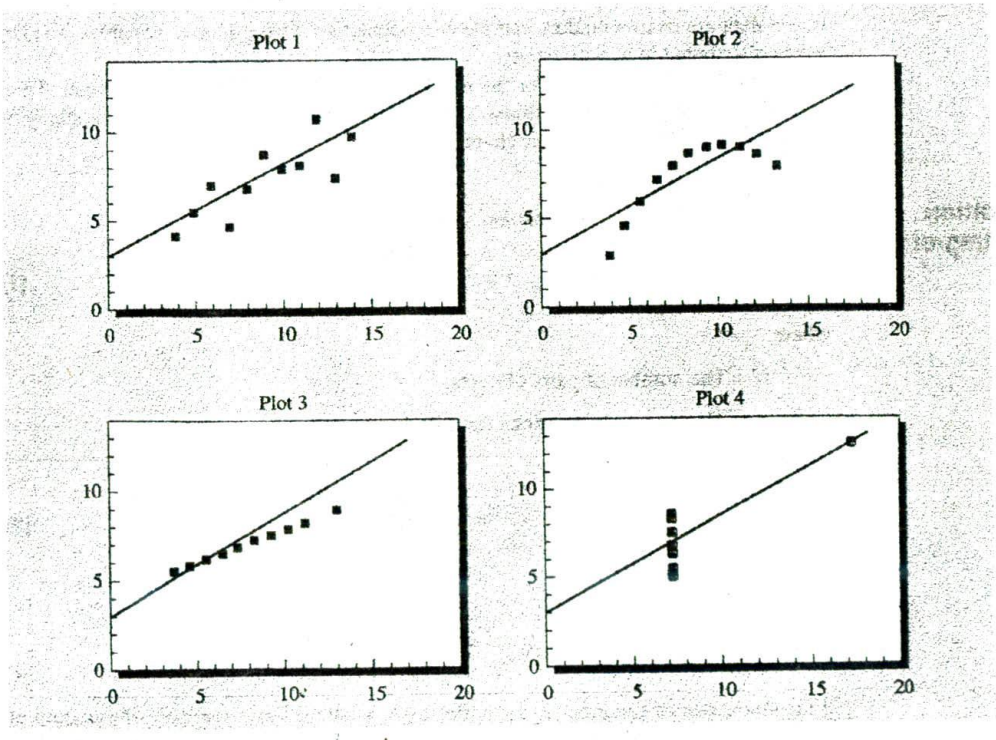
The shape of linear relationships is characterized by a straight line, whereas nonlinear relationships have curvilinear, parabolic, and compound curves representing their shapes. Pearson's r measures relationships in variables that are linearly related. It cannot distinguish linear from nonlinear data. Summary statistics alone do not reveal the appropriateness of the data for the model, as the following example illustrates.

One author constructed four small data sets possessing identical summary statistics but displaying strikingly different patterns.[2] Exhibit 18–3 contains these data and Exhibit 18–4 plots them. In Plot 1 of the figure, the variables are positively related.

EXHIBIT 18–3 Four Data Sets with the Same Summary Statistics

S_s	X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.10	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.10	4	5.39	19	12.50
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
Pearson's r	.81642		.81624		.81629		.81652	
r^2	.66654		.66624		.66632		.66671	
Adjusted r^2	.62949		.62916		.62925		.62967	
Standard error	1.23660		1.23721		1.23631		1.23570	

EXHIBIT 18–4 Different Scatterplots for the Same Summary Statistics



Their points follow a superimposed straight line through the data. This example is well suited to correlation analysis. In Plot 2, the data are curvilinear in relation to the line, and r is an inappropriate measure of their relationship. Plot 3 shows the presence of an influential point that changed a coefficient that would have otherwise been a perfect +1.0. The last plot displays constant values of x (similar to what you might find in an animal or quality-control experiment). One leverage point establishes the fitted line for these data.

We will return to these concepts and the process of drawing the line when we discuss regression. For now, comparing Plots 2 through 4 with Plot 1 suggests the importance of visually inspecting correlation data for underlying patterns. Careful analysts make scatterplots an integral part of the inspection and exploration of their data. Although small samples may be plotted by hand, statistical software packages save time and offer a variety of plotting procedures.

The Assumptions of *r*

Like other parametric techniques, correlation analysis makes certain assumptions about the data. Many of these assumptions are necessary to test hypotheses about the coefficient.

The first requirement for r is **linearity.** All of the examples in Exhibit 18–2 with the exception of $r = 0$ illustrate a relationship between variables that can be described by a straight line passing through the data cloud. When $r = 0$, no pattern is evident that could be described with a single line. Parenthetically, it is also possible to find coefficients of 0 where the variables are highly related but in a nonlinear form. As we have seen, plots make such findings evident.

The second assumption for correlation is a **bivariate normal distribution**—that is, the data are from a random sample of a population where the two variables are normally distributed in a joint manner.

Often these assumptions or the required measurement level cannot be met. Then the analyst should select a nonlinear or nonparametric measure of association, many of which are described later in this chapter.

Computation and Testing of *r*

The formula for calculating Pearson's r is

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(N - 1)s_x s_y} \tag{1}$$

where

N = The number of pairs of cases.

s_x, s_y = The standard deviations for X and Y.

Alternatively,

$$r = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}} \tag{2}$$

since

$$s_x = \sqrt{\frac{\Sigma x^2}{N}} \qquad s_y = \sqrt{\frac{\Sigma y^2}{N}}$$

If the numerator of equation (2) is divided by N, we have the *covariance*, the amount of deviation that the X and Y distributions have in common. With a positive covariance, the

EXHIBIT 18–5 Computation of Pearson's Product Moment Correlation

			Deviations from Means				
Corporation	Net Profits ($ mil.) X	Cash Flow ($ mil.) Y	$(X - \bar{X})x$	$(Y - \bar{Y})y$	xy	x^2	y^2
1	82.6	126.5	−93.84	−178.64	16763.58	8805.95	31912.25
2	89.0	191.2	−87.44	−113.94	9962.91	7645.75	12982.32
3	176.0	267.0	−0.44	−38.14	16.78	0.19	1454.66
4	82.3	137.1	−94.14	−168.04	15819.29	8862.34	28237.44
5	413.5	806.8	237.06	501.66	118923.52	56197.44	251602.56
6	18.1	35.2	158.34	−269.94	42742.30	25071.56	72867.60
7	337.3	425.5	160.86	120.36	19361.11	25875.94	14486.53
8	145.8	380.0	−30.64	74.86	−2293.71	938.81	5604.02
9	172.6	326.6	−3.84	21.36	−82.02	14.75	456.25
10	247.2	355.5	70.76	50.36	3563.47	5006.98	2536.13
	$\bar{X} = 176.44$	$\bar{Y} = 305.14$			$\Sigma xy = 224777.23$		
	$s_x = 216.59$	$s_y = 124.01$				$\Sigma x^2 = 138419.71$	
							$\Sigma y^2 = 422139.76$

variables move in unison; with a negative one, they move in opposition. When the covariance is 0, there is no relationship. The denominator for equation (2) represents the maximum potential variation that the two distributions share. Thus, correlation may be thought of as a ratio.

Exhibit 18–5 contains a random subsample of 10 firms of the Forbes 500 sample. The variables chosen to illustrate the computation of r are cash flow and net profits. Beneath each variable is its mean and standard deviation. In columns 4 and 5 we obtain the deviations of the X and Y values from their means, and in column 6 we find the product. Columns 7 and 8 are the squared deviation scores.

Substituting into the formula, we get

$$r = \frac{224777.23}{\sqrt{138419.71} * \sqrt{422139.76}} = .9298$$

In this subsample, net profits and cash flow are positively related and have a very high coefficient. As net profits increase, cash flow increases; the opposite is also true. Linearity of the variables may be examined with a scatterplot such as the one shown in Exhibit 18–6. The data points fall along a straight line.

Common Variance as an Explanation The amount of common variance in X (net profits) and Y (cash flow) may be summarized by r^2, the **coefficient of determination.** As Exhibit 18–7 shows, the overlap between the two variables is the proportion of their common or shared variance.

EXHIBIT 18–6 Plot of Forbes 500 Net Profits with Cash Flow

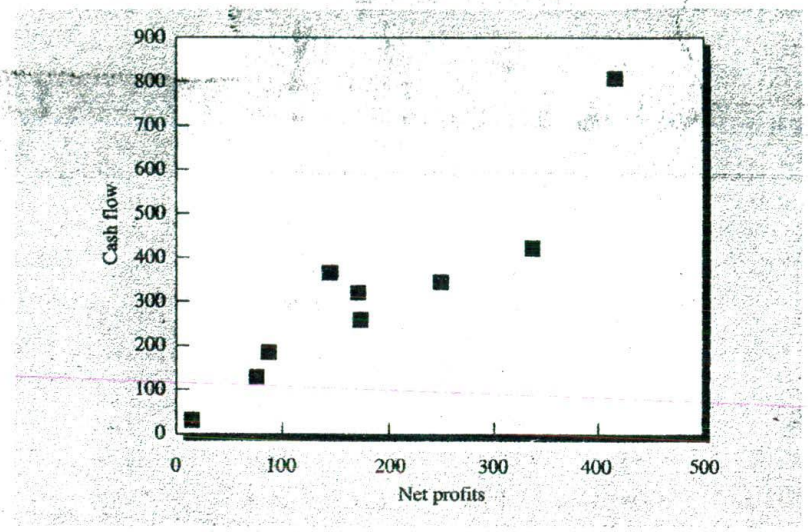
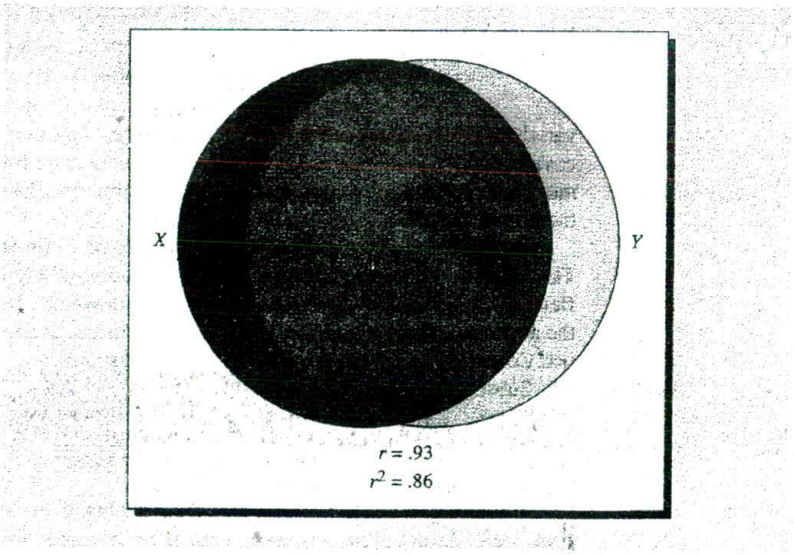


EXHIBIT 18–7 Diagram of Common Variance



The area of overlap represents the percentage of the total relationship accounted for by one variable or the other. So 86 percent of the variance in X is explained by Y, and vice versa.

Testing the Significance of *r* Is the coefficient representing the relationship between net profits and cash flow real, or does it occur by chance? This question tries to discover whether our r is a chance deviation from a population ρ of zero. In other situa-

tions, the researcher may wish to know if significant differences exist between two or more *r*s. In either case, *r*'s significance should be checked before *r* is used in other calculations or comparisons. For this test, we must have independent random samples from a bivariate normal distribution. Then the *Z* or *t*-test may be used for the null hypothesis, $p = 0$.

The formula for small samples is

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

where

$r = .93$

$n = 10$

Substituting into the equation, we calculate *t*:

$$t = \frac{.93}{\sqrt{\frac{1 - .86}{8}}} = 7.03$$

With $n - 2$ degrees of freedom, the statistical program calculates the value of t (7.03) at a probability less than .005 for the one-tailed alternative, H_A: $\rho > 0$. We reject the hypothesis that there is no linear relationship between net profits and cash flow in the population. The above statistic is appropriate when the null hypothesis states a correlation of 0. It should be used only for a one-tailed test.[3] However, it is often difficult to know in advance whether the variables are positively or negatively related, particularly when a computer removes our contact with the raw data. Software programs produce two-tailed tests for this eventuality. The observed significance level for a one-tailed test is half of the printed two-tailed version in most programs.

Correlation Matrix

A **correlation matrix** is a table used to display coefficients for more than two variables. Exhibit 18–8 shows the intercorrelations among six variables for the full Forbes 500 data set.[4]

EXHIBIT 18–8 Correlation Matrix for Forbes 500 Sample

	Assets ($ mil.)	Cash Flow ($ mil.)	Number Employed (thousands)	Market Value ($ mil.)	Net Profits ($ mil.)	Sales ($ mil.)
Assets	1.0000					
Cash flow	.3426	1.0000				
Employed	.3898	.8161	1.0000			
Market value	.3642	.9353	.8106	1.0000		
Net profits	.2747	.9537	.7467	.9101	1.0000	
Sales	.5921	.7990	.8831	.7485	.7261	1.0000

Notes: All coefficients are statistically significant, $p < .01$.

n = 100

It is conventional for a symmetrical matrix to report findings in the triangle below the diagonal. The diagonal contains coefficients of 1.00 that signify the relationship of each variable with itself. Journal articles and management reports often show matrices with coefficients at different probability levels. A symbol beside the coefficient keys the description of differences to a legend. The practice of reporting tests of the null hypothesis, $r = 0$, was followed in Exhibit 18–8.

Correlation matrices have utility beyond bivariate correlation studies. Interdependence among variables is a common characteristic of most multivariate techniques. Matrices form the basis for computation and understanding of the nature of relationships in multiple regression, discriminant analysis, factor analysis, and many others. Such applications call for variations on the standard matrix. Pooled within-groups covariance matrices average the separate covariances for several groups and array the results as coefficients. Total or overall correlation matrices treat coefficients as if they came from a single sample.

Interpretation of Correlations

You might want to review the nature of causation in Chapter 6.

A correlation coefficient of any magnitude or sign, whatever its statistical significance, does not imply causation. Increased net profits may cause an increase in market value, or improved satisfaction may cause improved performance in certain situations, but correlation provides no evidence of cause and effect. Several alternate explanations may be provided for correlation results:

- *X* causes *Y*.
- *Y* causes *X*.
- *X* and *Y* are activated by one or more other variables.
- *X* and *Y* influence each other reciprocally.

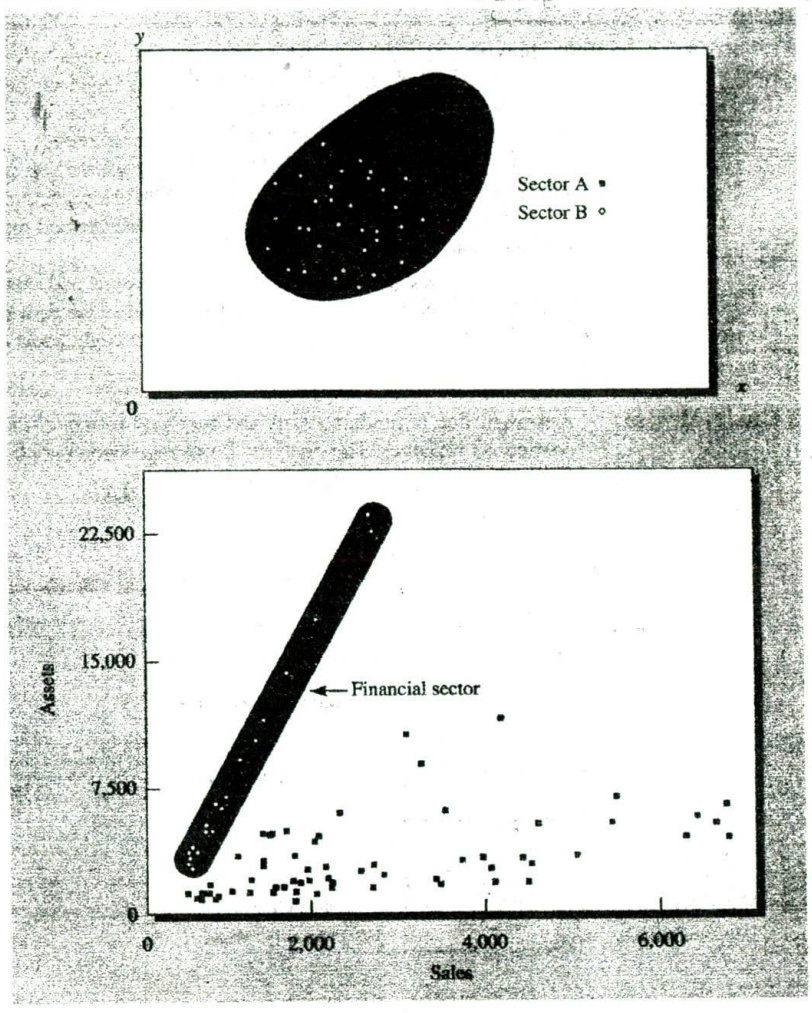
Ex post facto studies seldom possess sufficiently powerful designs to demonstrate which of these conditions could be true. By controlling variables under an experimental design, we may obtain more rigorous evidence of causality.

MANAGEMENT

Take care to avoid so-called **artifact correlations,** where distinct groups combine to give the impression of one. The upper panel of Exhibit 18–9 shows data from two business sectors. If all the data points for the *X* and *Y* variables are aggregated and a correlation is computed for a single group, a positive correlation results. Separate calculations for each sector (note that points for Sector A form a circle, as do points for Sector B) reveal *no* relationship between the *X* and *Y* variables. A second example shown in the lower panel contains a plot of data on assets and sales. We have enclosed and highlighted the data for the financial sector. This is shown as a narrow band enclosed by an ellipse. These companies score high on assets and low in sales—all are banks. When the data for banks are removed and treated separately, the correlation is nearly perfect (.99). When banks are returned to the sample and the correlation is recalculated, the overall relationship drops to the mid-.80s. In short, data hidden or nested within an aggregated set may present a radically different picture.

Another issue affecting interpretation of coefficients concerns practical significance. Even when a coefficient is statistically significant, it must be practically meaningful. In many relationships, other factors combine to make the coefficient's meaning misleading. For example, in nature we expect rainfall and the height of reservoirs to be positively correlated. But in states where water management and flood control mechanisms are complex, an apparently simple relationship may not hold. Techniques like partial and multiple correlation or multiple regression are helpful in sorting out confounding effects.

EXHIBIT 18–9 Artifact Correlations



With large samples, even exceedingly low coefficients can be statistically significant. This "significance" only reflects the likelihood of a linear relationship in the population. Should magnitudes less than .30 be reported when they are significant? It all depends. We might consider the correlations between variables such as cash flow, sales, market value, or net profits to be interesting revelations of a particular phenomenon whether they were high, moderate, or low. The nature of the study, the characteristics of the sample, or other reasons will be determining factors. But a coefficient is not remarkable simply because it is statistically significant.

By probing the evidence of direction, magnitude, statistical significance, and common variance together with the study's objectives and limitations, we reduce the chances of reporting trivial findings. Simultaneously, the communication of practical implications to the reader will be improved.

Bivariate Linear Regression[5]

In the previous section, we focused on relationships between variables. The product moment correlation was found to represent an index of the magnitude of the relationship, the sign governed the direction, and r^2 explained the common variance. Relationships also serve as a basis for estimation and prediction.

When we take the observed values of X to estimate or predict corresponding Y values, the process is called *simple prediction.*[6] When more than one X variable is used, the outcome is a function of multiple predictors. Simple and multiple predictions are made with a technique called **regression analysis.**

The similarities and differences of regression and correlation are summarized in Exhibit 18–10. Their relatedness would suggest that beneath many correlation problems is a regression analysis that could provide further insight about the relationship of Y with X.

The Basic Model

A straight line is fundamentally the best way to model the relationship between two continuous variables. The bivariate linear regression may be expressed as

$$Y = \beta_0 + \beta_1 X_i$$

EXHIBIT 18–10 Comparison of Bivariate Linear Correlation and Regression

	Correlation	Regression
Measurement level	Interval or ratio scale	Interval or ratio scale
Nature of variables	Both continuous, linearly related	Both continuous, linearly related
$X - Y$ relationship	X and Y are symmetric; $r_{xy} = r_{yx}$	Y is dependent, X is independent; regression of X on Y differs from Y on X
Correlation	The correlation of x and y produces an estimate of linear association based on sampling data	Correlation of $Y - X$ is the same as the correlation between the predicted values of Y and observed values of Y
Coefficient of determination	Explains common variance of X and Y	Proportion of variability of Y explained by its least-squares regression on X

where the value of the dependent variable Y is a linear function of the corresponding value of the independent variable X_i in the ith observation. The slope, β_1, and the Y intercept, β_0, are known as **regression coefficients.** The **slope,** β_1, is the change in Y for a one-unit change in X. It is sometimes called the "rise over run." This is defined by the formula

$$\beta_1 = \frac{\Delta Y}{\Delta X}$$

This is the ratio of change (Δ) in the rise of the line relative to the run or travel along the X axis. Exhibit 18–11 shows a few of the many possible slopes you may encounter.

The **intercept,** β_0, is the value for the linear function when it crosses the Y axis; it is the estimate of Y when $X = 0$. A formula for the intercept based on the mean scores of the X and Y variables is

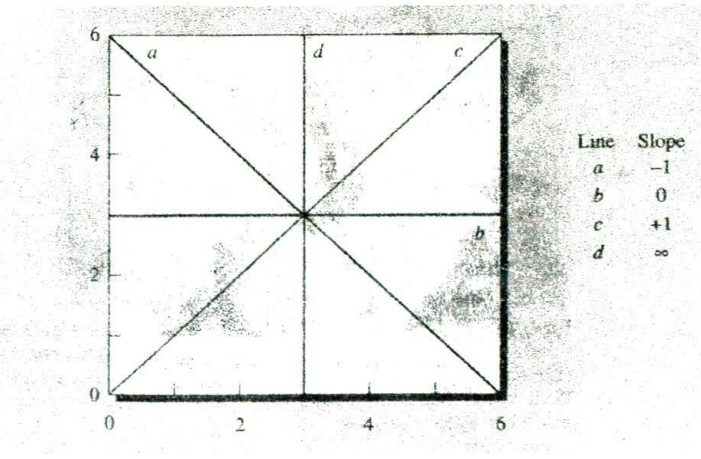
$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Concept Application

The price of investment-grade red wine is influenced in several ways, not the least of which is tasting. Tasting from the barrel is a major determinant of market *en primeur* or futures contracts, which represent about 60 percent of the harvest. After the wine rests for 18 to 24 months in oak casks, further tasting occurs, and the remaining stock is released.

Weather is widely regarded as responsible for pronouncements about wine quality. A Princeton economist has elaborated on that notion. He suggested that just a few facts about local weather conditions may be better predictors of vintage French red wines than the most refined palates and noses.[7] The regression model developed predicts an auction price index for about 80 wines from winter and harvest rainfall amounts and average growing-season temperatures. Interestingly, the calculations suggested that the 1989 Bordeaux would be one of the best since 1893. The "guardians of tradition" reacted hysterically to these methods yet agreed with the conclusion.

EXHIBIT 18–11 Examples of Different Slopes



Our first example will use one predictor with highly simplified data. Let X represent the average growing-season temperature in degrees Celsius and Y the price of a 12-bottle case in French francs. The data appear below.

X Average Temperature Celsius	Y Price per Case (FF)
12	2000
16	3000
20	4000
24	5000
$\bar{X} = 18$	$\bar{Y} = 3500$

The plotted data in Exhibit 18–12 show a linear relationship between the pairs of points and a perfect positive correlation, $r_{yx} = 1.0$. The slope of the line is calculated:

$$\beta_1 = \frac{Y_i - Y_j}{X_i - X_j} = \frac{4000 - 3000}{20 - 16} = \frac{1000}{4} = 250$$

where the X_iY_i values are the data points (20, 4000) and X_jY_j are points (16, 3000). The intercept β_0 is –1000, the point at which $X = 0$ in this plot. This area is off the graph and appears in an insert on the figure.

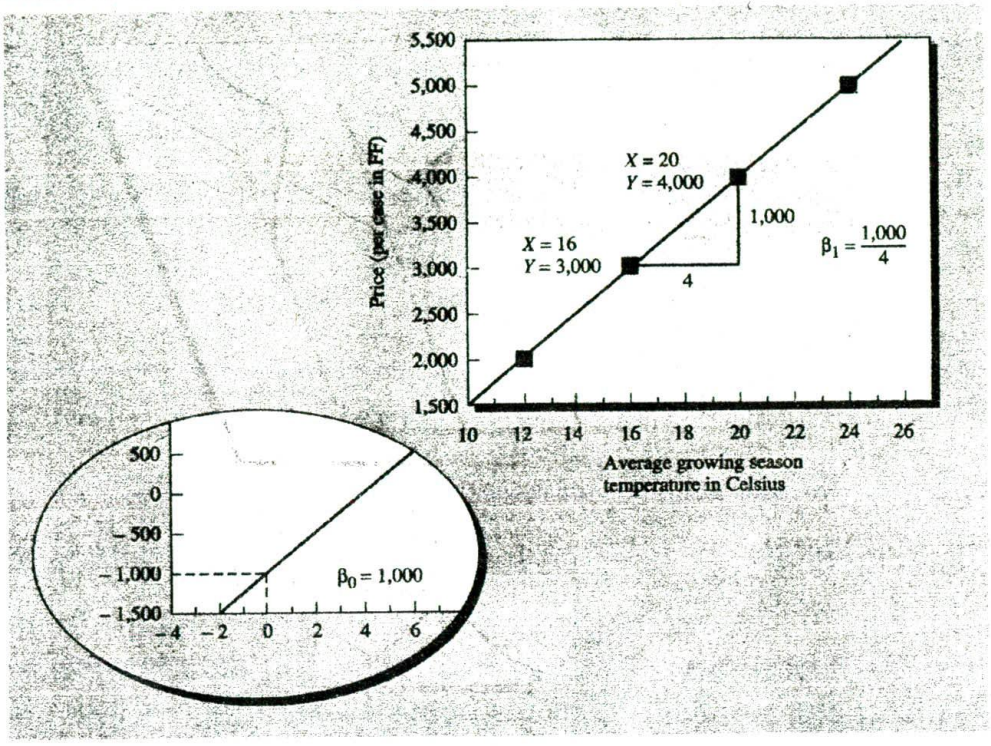
$$\beta_0 = \bar{Y} - \beta_1\bar{X} = 3500 - 250\ (18) = -1000$$

Substituting into the formula, we have the simple regression equation

$$Y = -1000 + 250\ X_i$$

France's Bordeaux Business School offers a master of business administration in the wine sector. Surprised? Business schools throughout Europe are increasingly tailoring their programs with innovative degrees that respond to the changing environment of business. In addition to wine, MBA specializations focus on the music industry, luxury brands, sports management, agribusiness, e-business, consulting, and public-sector specialities.
www.iht.com

EXHIBIT 18–12 Plot of Wine Price by Average Growing Temperature



We could now predict that a warm growing season with 25.5°C temperature would bring a case price of 5375 French francs. $\hat{Y}$ (called *Y*-hat) is the predicted value of *Y*.

$$\hat{Y} = -1000 + 250(25.5) = 5375$$

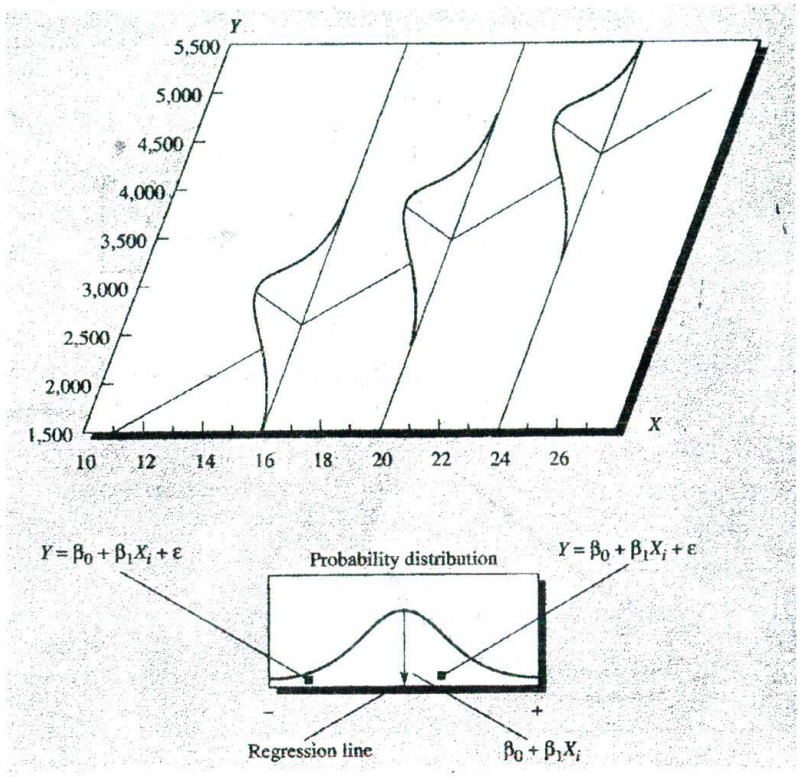
Unfortunately, one rarely comes across a data set composed of four paired values, a perfect correlation, and an easily drawn line. A model based on such data is *deterministic* in that for any value of *X*, there is only one possible corresponding value of *Y*. It is more likely that we will collect data where the values of *Y* vary for each *X* value. Considering Exhibit 18–13, we should expect a distribution of price values for the temperature $X = 16$, another for $X = 20$, and another for *each* value of *X*. The means of these *Y* distributions will also vary in some systematic way with *X*. These variabilities lead us to construct a *probabilistic* model that also uses a linear function.[8] This function is written

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_1$$

where ε symbolizes the deviation of the *i*th observation from the mean, $\beta_0 + \beta_1 X_i$.

As shown in Exhibit 18–13, the actual values of *Y* may be found above or below the regression line represented by the mean value of *Y* ($\beta_0 + \beta_1 X_i$) for a particular value of *X*. These deviations are the error in fitting the line and are often called the **error term.**

EXHIBIT 18–13 Distribution of *Y* for Observations of *X*



Method of Least Squares

Exhibit 18–14 contains a new data set for the wine price example. Our prediction of *Y* from *X* must now account for the fact that the *X* and *Y* pairs do not fall neatly along the line. Actually, the relationship could be summarized by several lines. Exhibit 18–15 suggests a few alternatives based on visual inspection—all of which produce errors, or vertical distances from the observed values to the line. The **method of least squares** allows us to find a regression line, or line of best fit, which will keep these errors to a minimum. It uses the criterion of minimizing the total squared errors of estimate. When we predict values of *Y* for each X_i, the difference between the actual Y_i and the predicted $\hat{Y}$ is the error. This error is squared and then summed. The line of best fit is the one that minimizes the total squared errors of prediction.[9]

$$\sum_{i=1}^{n} e_i^2 \text{ minimized}$$

Regression coefficients β_0 and β_1 are used to find the least-squares solution. They are computed as follows:

$$\beta_1 = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Substituting data from Exhibit 18–14 into both formulas, we get

$$\beta_1 = \frac{748633.5 - \frac{(196.1)(35988)}{10}}{4043.77 - \frac{(196.1)^2}{10}} = 216.439$$

$$\hat{\beta}_0 = 3598.8 - (216.439)(19.61) = -645.569$$

The predictive equation is now $\hat{Y} = -645.57 + 216.44\,X_i$.

Drawing the Regression Line Before drawing the regression line, we select two values of X to compute. Using values 13 and 24 for X_i, the points are

$$\hat{Y} = -645.57 + 216.44(13) = 2168.15$$

$$\hat{Y} = 645.57 + 216.44(24) = 4548.99$$

Comparing the line drawn in Exhibit 18–16 to the trial lines in Exhibit 18–15, one can readily see the success of the least-squares method in minimizing the error of prediction.

Residuals We now turn our attention to the plot of standardized residuals in Exhibit 18–17. A **residual** is what remains after the line is fit or $(Y_i - \hat{Y}_i)$. When standardized, residuals are comparable to Z scores with a mean of 0 and a standard deviation of 1. In this plot, the standardized residuals should fall between 2 and –2, be randomly distributed about zero, and show no discernible pattern. All these conditions say the model is applied correctly.

EXHIBIT 18–14 Data for Wine Price Study

	Y Price (FF)	X Temperature (C°)	XY	Y^2	X^2
1	1813.00	11.80	21393.40	3286969.00	139.24
2	2558.00	15.70	40160.60	6543364.00	246.49
3	2628.00	14.00	36792.00	6906384.00	196.00
4	3217.00	22.90	73669.30	10349089.00	524.41
5	3228.00	20.00	64560.00	10419984.00	400.00
6	3629.00	20.10	72942.90	13169641.00	404.01
7	3886.00	17.90	69559.40	15100996.00	320.41
8	4897.00	23.40	114589.80	23980609.00	547.56
9	4933.00	24.60	121351.80	24334489.00	605.16
10	5199.00	25.70	133614.30	27029601.00	660.49
Σ	35988.00	196.10	748633.50	141121126.00	4043.77
Mean	3598.80	19.61			
s	1135.66	4.69			
Sum of squares (SS)	11607511.59	198.25	42908.82		

EXHIBIT 18–15 Scatterplot and Possible Regression Lines Based on Visual Inspection: Wine Price Study

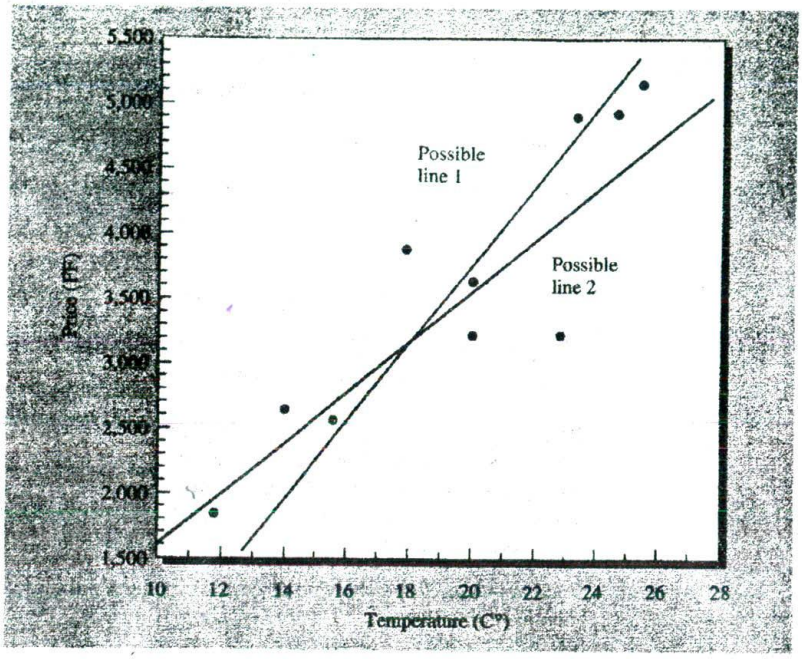
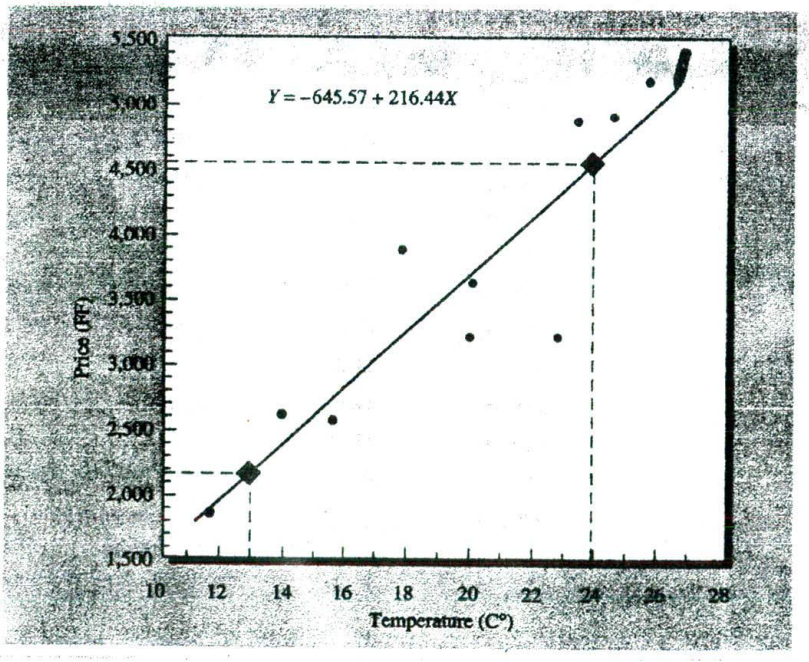


EXHIBIT 18–16 Drawing the Least-Squares Line: Wine Price Study



SNAPSHOT

What's a Business Education without Wine?

What do Harvard, Yale, UCLA, Columbia, the Kellogg School of Management at Northwestern, Pennsylvania's Wharton Business School, and Berkeley's Hass School of Business have in common? They are among a growing number of B-schools where wine clubs have flourished. Some have even added wine education to the business curriculum.

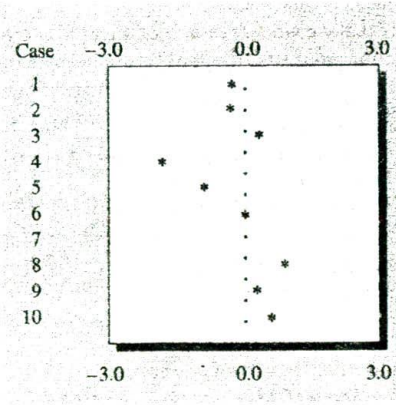
While medical research has shown moderate drinking to reduce the risk of heart disease, that's not the appeal for students who believe that it can be an effective tool for shaping positive business relationships. Brian Scanlon of Harvard's Wine & Cuisine Society summed it up this way: "Wine knowledge is an indispensable skill in today's business environment. If you're at a crucial business dinner and you want to pick the perfect wine to create the right atmosphere, you need to know the vintages, the regions and the best winemakers."

Vineyard owners couldn't be more supportive. Jack Cakebread of Cakebread Cellars, on a promotional tour at business schools around the country, reported the relationship between age and visitation frequency at their tasting room. Almost 70 percent of visitors are in their 20s and 30s. Although wine industry research forecasts a drop in wine enthusiasm for "Generation X," the future corporate executives represent a radically different segment.

David Mogridge is on a student team at Berkeley that brings in lecturers on a wide range of topics like growing, shipping, legal issues, branding, and strategy. In an interview with Eric Zelko of *Wine Spectator*, Mogridge said playfully, "When I think about it, everything I learned in business school, I learned in wine class."

www.winespectator.com

EXHIBIT 18–17 Plot of Standardized Residuals: Wine Price Study



Case	Y Price	Predicted Price	Residual
1	1,813	1,908.4112	−95.4112
2	2,558	2,752.5234	−194.5234
3	2,628	2,384.5771	243.4229
4	3,217	4,310.8844	−1,093.8844
5	3,228	3,683.2112	−455.2112
6	3,629	3,704.8551	−75.8551
7	3,886	3,228.6893	657.3107
8	4,897	4,419.1039	477.8961
9	4,933	4,678.8307	254.1693
10	5,199	4,916.9137	282.0863

In our example, we have one residual at −2.2, a random distribution about zero, and few indications of a sequential pattern. It is important to apply other diagnostics to verify that the regression assumptions are met. Various software programs provide plots and other checks of normality, linearity, equality of variance, and independence of error.[10]

Predictions

If we wanted to predict the price of a case of investment-grade red wine for a growing season that averages 21°C, our prediction would be

$$\hat{Y} = -645.57 + 216.44(21) = 3899.67$$

This is a *point prediction* of Y and should be corrected for greater precision. As with other confidence estimates, we establish the degree of confidence desired and substitute into the formula

$$\hat{Y} \pm t_{\alpha/2^s}\sqrt{1 + \frac{1}{10} + \frac{(X - \bar{X})^2}{SS_x}}$$

where

$t_{\alpha/2}$ = The two-tailed critical value for t at the desired level (95 percent in this example).

s = The standard error of estimate (also the square root of the mean square error from the analysis of variance of the regression model) (see Exhibit 18–20).

SS_x = The sum of squares for X (Exhibit 18–14).

$$3899.67 \pm (2.306)(538.559)\sqrt{1 + \frac{1}{10} + \frac{(21 - 19.61)^2}{198.25}}$$

$$3899.67 \pm 1308.29$$

We are 95 percent confident of our prediction that a case of investment-quality French red wine grown in a particular year at 21°C average temperatures will be initially priced at 3899.67 ± 1308.29, or from approximately 2591 to 5208 FF. The comparatively large band width results from the amount of error in the model (reflected by r^2), some peculiarities in the Y values, and the use of a single predictor.

It is more likely that we would want to predict the average price of *all* cases grown at 21°C. This prediction would use the same basic formula omitting the first digit (the 1) under the radical. A narrower *confidence* band is the result since the average of all Y values is being predicted from a given X. In our example, the confidence interval for 95 percent is 3899.67 ± 411.42, or from 3488 to 4311 FF.

The predictor we selected, 21°C, was close to the mean of X (19.61). Because the **prediction and confidence bands** are shaped like a bow tie, predictors farther from the mean have larger band widths. For example, X values of 15, 20, and 25 produce confidence bands of ±565, ±397, and ±617, respectively. This is illustrated in Exhibit 18–18. The farther one's selected predictor is from X, the wider is the prediction interval.

Testing the Goodness of Fit

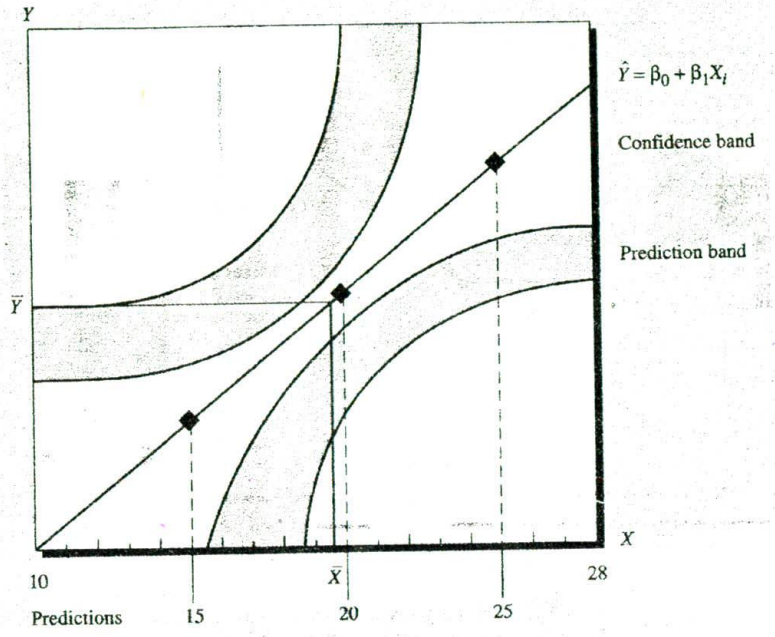
With the regression line plotted and a few illustrative predictions, we should now gather some evidence of **goodness of fit**—how well the model fits the data. The most important test in bivariate linear regression is whether the slope, β_1, is equal to zero.[11] We have already observed a slope of zero in Exhibit 18–11, line *b*. Zero slopes result from various conditions:

- Y is completely unrelated to X, and no systematic pattern is evident.
- There are constant values of Y for every value of X.
- The data are related but represented by a nonlinear function.

The *t*-Test To test whether $\beta_1 = 0$, we use a two-tailed test (since the actual relationship is positive, negative, or zero). The test follows the t distribution for $n - 2$ degrees of freedom.

$$t = \frac{b_1}{s(b_1)} = \frac{216.439}{34.249} = 5.659$$

EXHIBIT 18–18 Prediction and Confidence Bands Based on Proximity to *X*



where

b_1 was previously defined as the slope β_1.

$s(b_1)$ is the standard error of β_1.[12]

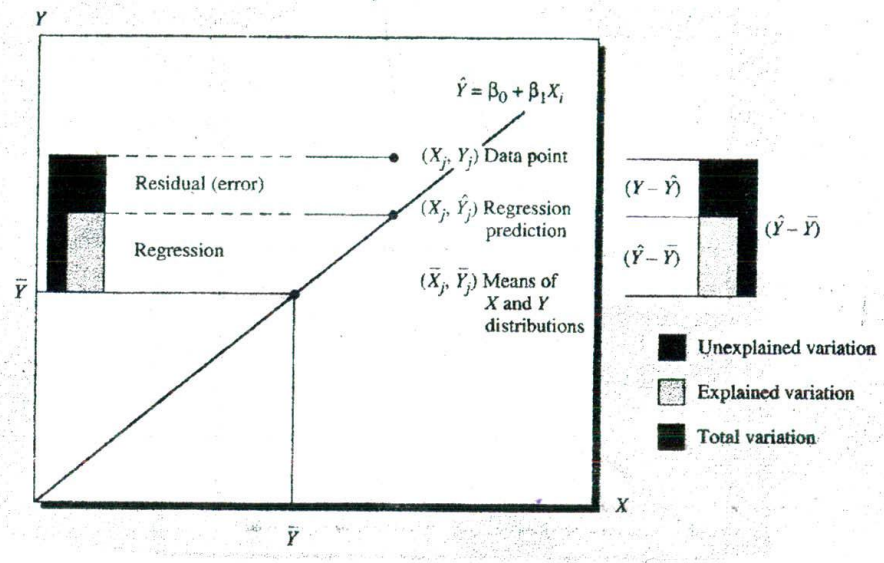
We reject the null, $\beta_1 = 0$, because the calculated t is greater than any t value for 8 degrees of freedom and $\alpha = .01$.

The *F* Test Computer printouts generally contain an analysis of variance (ANOVA) table with an F test of the regression model. In bivariate regression, t and F tests produce the same results since t^2 is equal to F. In multiple regression, the F test has an overall role for the model, and each of the independent variables is evaluated with a separate t-test. From the last chapter, recall that ANOVA partitions variance into component parts. For regression, it comprises explained deviations, $\hat{Y} - \bar{Y}$, and unexplained deviations, $Y - \hat{Y}$. Together they constitute the total deviation, $Y - \bar{Y}$. This is shown graphically in Exhibit 18–19. These sources of deviation are squared for all observations and summed across the data points.

In Exhibit 18–20, we develop this concept sequentially concluding with the F test of the regression model for the wine data. Based on the results presented in that table, we find statistical evidence of a linear relationship between variables. The alternative hypothesis, $r^2 \neq 0$, is accepted with $F = 32.02$, d.f., $(1,8)$, $p < .005$. The null hypothesis for the F test had the same effect as $\beta_1 = 0$ since we could select either test.

Coefficient of Determination In predicting the values of Y without any knowledge of X, our best estimate would be $\bar{Y}$, its mean. Each predicted value that does not

EXHIBIT 18–19 Components of Variation



fall on Y contributes to an error of estimate, $(Y - \bar{Y})$. The total squared error for several predictions would be $\Sigma(Y_i - \bar{Y})^2$. By introducing known values of X into a regression equation, we attempt to reduce this error even further. Naturally, this is an improvement over using $\bar{Y}$, and the result is $(\hat{Y} - \bar{Y})$. The total improvement based on several estimates is $\Sigma(\hat{Y}_i - \bar{Y})^2$, the amount of variation explained by the relationship between X and Y in the regression. Based on the formula, the *coefficient of determination* is the ratio of the line of best fit's error over that incurred by using Y. One purpose of testing, then, is to discover whether the regression equation is a more effective predictive device than the mean of the dependent variable.

As in correlation, the coefficient of determination is symbolized by r^2.[13] It has several purposes. As an index of fit, it is interpreted as the total proportion of variance in Y explained by X. As a measure of linear relationship, it tells us how well the regression line fits the data. It is also an important indicator of the predictive accuracy of the equation. Typically, we would like to have an r^2 that explains 80 percent or more of the variation. Lower than that, predictive accuracy begins to fall off. The coefficient of determination, r^2, is calculated like this:

$$r^2 = \frac{\sum_{i=1}^{n} (\hat{Y} - \bar{Y})^2}{\sum_{i=1}^{n} (Y - \bar{Y})^2} = \frac{SS_r}{SS_e} = 1 - \frac{SS_e}{SS_t}$$

For the wine price study, r^2 was found by using the data from the bottom of Exhibit 18–20.

$$r^2 = 1 - \frac{2320368.49}{11607511.60} = .80$$

Eighty percent of the variance in price may be explained by growing-season temperatures. With actual data and multiple predictors, our results would improve.

EXHIBIT 18–20 Progressive Application of Partitioned Variance Concept

General Concept

$(\hat{Y} - \bar{Y})$	+	$(Y - \hat{Y})$	=	$(Y - \bar{Y})$
Explained Variation (the regression relationship between X and Y)		Unexplained Variation (cannot be explained by the regression relationship)		Total Variation

ANOVA Application

$\sum_{i=1}^{n} (\hat{Y} - \bar{Y})^2$	$\sum_{i=1}^{n} (Y - \hat{Y})^2$	$\sum_{i=1}^{n} (Y - \bar{Y})^2$
SS_r Sum of Squares Regression	SS_e Sum of Squares Error	SS_t Sum of Squares Total

Contents of Summary Table

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio
Regression	1	SS_r	$MS_r = \frac{SS_r}{1}$	MS_r
Error	$n - 2$	SS_e	$MS_e = \frac{SS_e}{n-2}$	$\overline{MS_e}$
Total		SS_t		

ANOVA Summary Table for [illegible]

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio
Regression	1	9,287,143.11	9,287,143.11	32.02
Residual (error)	8	2,320368.49	290,046.06	
Total		11,607,511.60		

Significance of $F = .0005$

Nonparametric Measures of Association[14]

Measures for Nominal Data

Nominal measures are used to assess the strength of relationships in cross-classification tables. They are often used with chi-square or may be used separately. In this section, we provide examples of three statistics based on chi-square and two that follow the proportional reduction in error approach.

There is no fully satisfactory all-purpose measure for categorical data. Some are adversely affected by table shape and number of cells; others are sensitive to sample size or marginals. It is perturbing to find similar statistics reporting different coefficients for the same data. This occurs because of a statistic's particular sensitivity or the way it was devised.

Technically, we would like to find two characteristics with nominal measures:

- When there is no relationship at all, the coefficient should be 0.
- When there is a complete dependency, the coefficient should display unity or 1.

This does not always happen. In addition to the sensitivity problem, analysts should be alerted to the need for careful selection of tests.

You may wish to review our discussion of chi-square in Chapter 17.

Chi-Square-Based Measures Exhibit 18–21 reports a 2 × 2 variation of the Containers Inc. shipping study on smoking and job-related accidents introduced in Chapter 17. In this example, the observed significance level is less than the testing level (α =

EXHIBIT 18–21 Chi-Square-Based Measures of Association

		On-the-Job Accident		
	Count	Yes	No	Row Total
Smoker	Yes	21	10	31
	No	13	22	35
	Column Total	34	32	66

Chi-Square	*Value*	*d.f.*	*Significance*
Pearson	6.16257	1	.01305
Community correction	4.99836	1	.02537

Minimal expected frequency 15.030

Statistic	*Value*	*Approximate Significance*
Phi	.30557	.01305*
Cramer's *V*	.30557	.01305*
Contingency coefficient *C*	.29223	.01305*

*Pearson chi-square probability.

.05), and the null hypothesis is rejected. A correction to chi-square is provided. We now turn to measures of association to detect the strength of the relationship. Notice that the exhibit also provides an approximate significance of the coefficient based on the chi-square distribution. This is a test of the null hypothesis that no relationship exists between the variables of accidents and smoking.

The first **chi-square-based measure** is applied to smoking and on-the-job accidents. It is called **phi** (ϕ). Phi ranges from 0 to +1.0 and attempts to correct χ^2 proportionately to N. Phi is best employed with 2×2 tables like this one since its coefficient can exceed +1.0 when applied to larger tables. Phi is calculated:

$$\phi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{6.616257}{66}} = .3056$$

Phi's coefficient shows a moderate relationship between smoking and job-related accidents. There is no suggestion in this interpretation that one variable causes the other, nor is there an indication of the direction of the relationship.

Cramer's V is a modification of phi for larger tables and has a range up to 1.0 for tables of any shape. It is calculated like this:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} = \sqrt{\frac{6.616257}{66(1)}} = .3056$$

where

k = the lesser number of rows or columns.

In Exhibit 18–21, the coefficient is the same as phi.

The **contingency coefficient** C is reported last. It is not comparable to other measures and has a different upper limit for various table sizes. The upper limits are determined as

$$\sqrt{\frac{k-1}{k}}$$

where

k = the number of columns.

For a 2×2 table, the upper limit is .71; for a 3×3, .82; and for a 4×4, .87. Although this statistic operates well with tables having the same number of rows as columns, its upper-limit restriction is not consistent with a criterion of good association measurement. C is calculated as

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{6.616257}{6.61625} + 66} = .2922$$

The chief advantage of C is its ability to accommodate data in almost every form: skewed or normal, discrete or continuous, and nominal or ordinal.

Proportional Reduction in Error **Proportional reduction in error (PRE)** statistics are the second type used with contingency tables. Lambda and tau are the examples discussed here. The coefficient **lambda** (λ) is based on how well the frequencies of one nominal variable offer predictive evidence about the frequencies of another. Lambda is asymmetrical—allowing calculation for the direction of prediction—and symmetrical, predicting row and column variables equally.

EXHIBIT 18–22 Proportional Reduction in Error Measures

What is your opinion about capping executives' salaries?

	Count Row Pct.	Favor	Do not Favor	Row Total
Occupational Class	Managerial	90 82.0	20 18.0	110
	White collar	60 43.0	80 57.0	140
	Blue collar	30 20.0	120 80.0	150
	Column Total	180 45.0%	220 55.0%	400 100.0%

Chi-Square	*Value*	*d.f.*	*Significance*
Pearson	98.38646	2	.00000
Likelihood ratio	104.96542	2	.00000

Minimum expected frequency 49.500

Statistic	*Value*	*ASEI*	*T Value*	*Approximate Significance*
Lambda:				
symmetric	.30233	.03955	6.77902	
with occupation dependent	.24000	.03820	5.69495	
with opinion dependent	.38889	.04555	7.08010	
Goodman & Kruskal tau:				
with occupation dependent	.11669	.02076		.00000*
with opinion dependent	.24597	.03979		.00000*

*Based on chi-square approximation.

The computation of lambda is straightforward. In Exhibit 18–22, we have results from an opinion survey with a sample of 400 shareholders. Only 180 out of 400 (45 percent) favor capping executives' salaries; 220 (55 percent) do not favor it. With this information alone, if asked to predict the opinions of an individual in the sample, we would achieve the best prediction record by always choosing the modal category. Here it is "do not favor." By doing so, however, we would be wrong 180 out of 400 times. The probability estimate for an incorrect classification is .45, $P(1) = (1 - .55)$.

Now suppose we have prior information about the respondents' occupational status and are asked to predict opinion. Would it improve predictive ability? Yes, we would make the predictions by summing the probabilities of all cells that are not the modal value for their rows (for example, cell [2, 1] is 20/400 or .05):

$$P(2) = \text{cell } (1, 2)\ .05 + \text{cell } (2, 1)\ .15 + \text{cell } (3, 1)\ .075 = .275$$

Lambda is then calculated:

$$\lambda = \frac{P(1) - P(2)}{P(1)} = \frac{.45 - .275}{45} = .3889$$

Note that the asymmetric lambda in Exhibit 18–22, where opinion is the dependent variable, reflects this computation. As a result of knowing the respondents' occupational classification, we improve our prediction by 39 percent. If we wish to predict occupational classification from opinion instead of the opposite, a λ of .24 would be secured. This means that 24 percent of the error in predicting occupational class is eliminated by knowledge of opinion on the executives' salary question. Lambda varies between 0 and 1, corresponding with no ability to eliminate errors to elimination of all errors of prediction.

Goodman and Kruskal's **tau** (τ) uses table marginals to reduce prediction errors. In predicting opinion on executives' salaries without any knowledge of occupational class, we would expect a 50.5 percent correct classification and a 49.5 percent probability of error. These are based on the column marginal percentages in Exhibit 18–22.

Column Marginal		Column Percent		Correct Cases
180	*	45	=	81
220	*	55	=	121
Total correct classification				202

Correct classification of the opinion variable = $.505 = \frac{202}{400}$

Probability of error, $P(1) = (1 - .505) = .495$

When additional knowledge of occupational class is used, information for correct classification of the opinion variable is improved to 62.7 percent with a 37.3 percent probability of error. This is obtained by using the cell counts and marginals for occupational class (refer to Exhibit 18–22), as shown below:

Row 1	$\left(\frac{90}{110}\right)90 + \left(\frac{20}{110}\right)20$	=	73.6364 + 3.6364	=	77.2727
Row 2	$\left(\frac{60}{140}\right)60 + \left(\frac{80}{140}\right)80$	=	25.7143 + 45.7142	=	71.4286
Row 3	$\left(\frac{30}{150}\right)30 + \left(\frac{120}{150}\right)120$	=	6.0 + 96.0	=	102.0000
Total correct classification (with additional information on occupational class)					250.7013

Correct classification of opinion variable = $.627 = \frac{250.7}{400}$

Probability of error, $P(2) = (1 - .627) = .373$

Tau is then computed like this:

$$\tau \frac{P(1) - P(2)}{P(1)} = \frac{.495 - .373}{.495} = .246$$

SNAPSHOT

Speedpass Is McD-lment

Lots of people will be waving while visiting McDonald's in Chicago. And it won't be because they want to attract attention or be overly friendly. Instead, Chicagoland McDonald's will be expanding a test of a cashless payment system first tested by McDonald's franchises in New York and Southern California and tested earlier in 2001 by nine Chicago McDonald's restaurant owners. The cashless payment system activates a charge on a credit card for any Big Mac or Egg McMuffin ordered. The system, called Speedpass, is activated when a customer waves a Speedpass card at a card reader located in either the drive-thru or inside at the checkout counter. The Speedpass system was originally introduced by Exxon Mobil Corp. at its Mobil gas stations. Similar systems have been tested by Taco Bell and KFC. How should this study be designed to measure the effectiveness of cashless payment systems? What relationships do you expect to find? Will they require parametric or nonparametric measures of association?

www.speedpass.com

Exhibit 18–22 shows that the information about occupational class has reduced error in predicting opinion to approximately 25 percent. The table also contains information on the test of the null hypothesis that tau = 0 with an approximate observed significance level and asymptotic error (for developing confidence intervals). Based on the small observed significance level, we would conclude that tau is significantly different from a coefficient of 0 and that there is an association between opinion on executives' salaries and occupational class in the population from which the sample was selected. We can also establish the confidence level for the coefficient at the 95 percent level as approximately .25 ± .04.

Measures for Ordinal Data

When data require **ordinal measures,** there are several statistical alternatives. In this section we will illustrate:

- Gamma.
- Kendall's tau *b* and tau *c*.
- Somers's *d*.
- Spearman's rho.

All but Spearman's rank-order correlation are based on the concept of concordant and discordant pairs. None of these statistics require the assumption of a bivariate normal distribution, yet by incorporating order, most produce a range from −1.0 (a perfect negative relationship) to +1.0 (a perfect positive one). Within this range, a coefficient with a larger magnitude (absolute value of the measure) is interpreted as having a stronger relationship. These characteristics allow the analyst to interpret both the direction and the strength of the relationship.

Exhibit 18–23 presents data for 70 managerial employees of KeyDesign, a large industrial design firm. All 70 employees have been evaluated for coronary risk by the firm's health insurer. The management levels are ranked, as are the fitness assessments by the physicians. If we were to use a nominal measure of association with this data (such as Cramer's *V*), the computed value of the statistic would be positive since order is not present in nominal data. But using ordinal measures of association reveals the actual nature of the relationship. In this example, all coefficients have negative signs.

The information in the exhibit has been arranged so the number of concordant and discordant pairs of individual observations may be calculated. When a subject that

EXHIBIT 18–23 Tabled Ranks for Management and Fitness Levels at Key Design

		Management Level			
	Count	Lower	Middle	Upper	
Fitness	High	14	4	2	20
	Moderate	18	6	2	26
	Low	2	6	16	24
		34	16	20	70

Statistic	*Value**
Gamma	–.70242
Kendall's tau *b*	–.51279
Kendall's tau *c*	–.49714
Somers's *d*	
Symmetric	–.51263
With fitness dependent	–.52591
With management-level dependent	–.50000

*The *t* value for each coefficient is –5.86451.

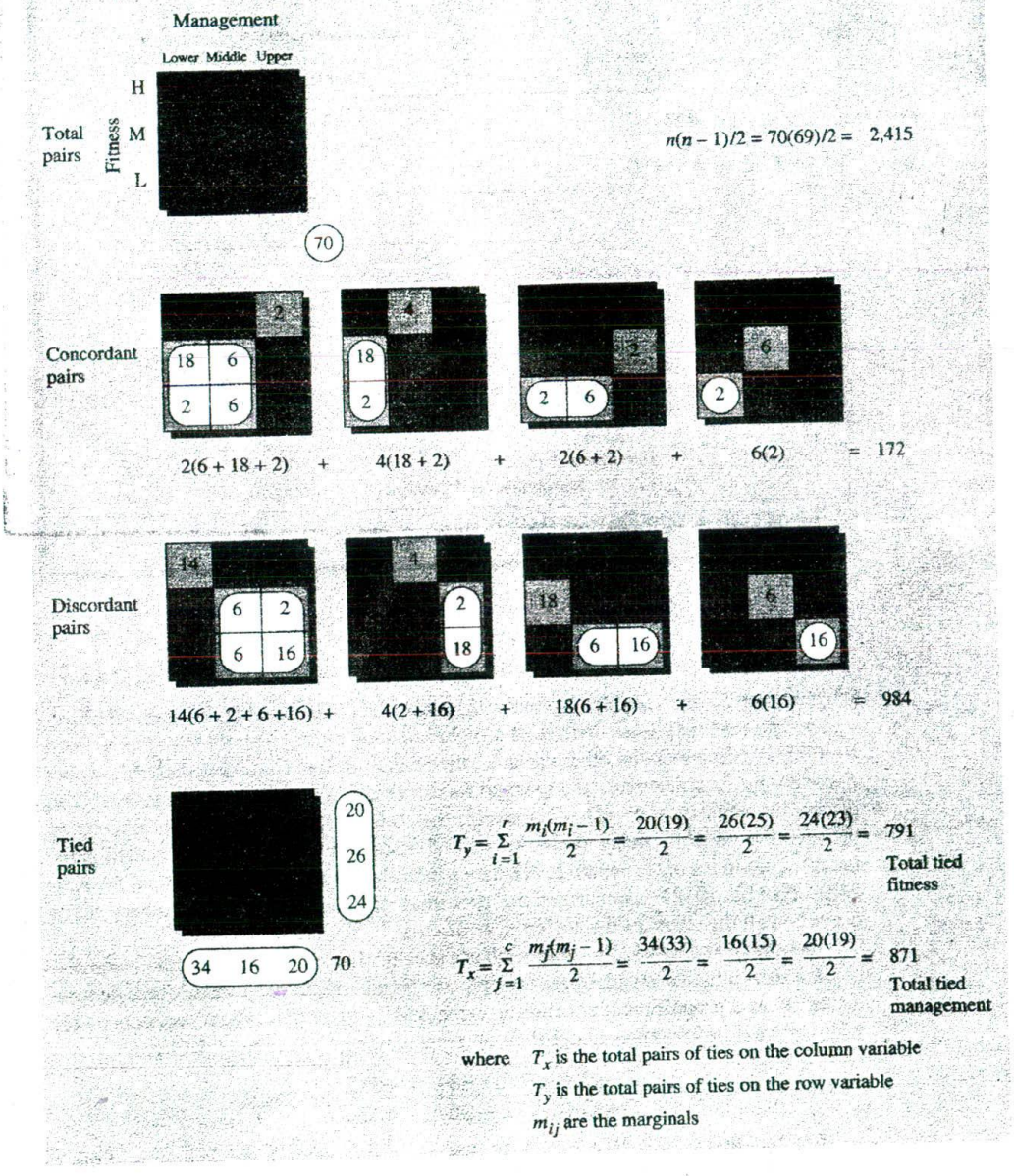
ranks higher on one variable also ranks higher on the other variable, the pairs of observations are said to be **concordant.** If a higher ranking on one variable is accompanied by a lower ranking on the other variable, the pairs of observations are **discordant.** Let P stand for concordant pairs and Q stand for discordant. When concordant pairs exceed discordant pairs in a $P - Q$ relationship, the statistic reports a positive association between the variables under study. As discordant pairs increase over concordant pairs, the association becomes negative. A balance indicates no relationship between the variables. Exhibit 18–24 summarizes the procedure for calculating the summary terms needed in all the statistics we are about to discuss.[15]

Goodman and Kruskal's **gamma** (γ) is a statistic that compares concordant and discordant pairs and then standardizes the outcome by maximizing the value of the denominator. It has a proportional reduction in error (PRE) interpretation that connects nicely with what we already know about PRE nominal measures. Gamma is defined as

$$\gamma = \frac{P - Q}{P + Q} = \frac{172 - 984}{172 + 984} = \frac{-812}{1156} = -.7024$$

For the fitness data, we conclude that as management level increases, fitness decreases. This is immediately apparent from the larger number of discordant pairs. A more precise explanation for gamma takes its absolute value (ignoring the sign) and relates it to PRE. Hypothetically, if one was trying to predict whether the pairs were concordant or discordant, one might flip a coin and classify the outcome. A better way is to make the prediction based on the preponderance of concordance or discordance;

EXHIBIT 18–24 Calculation of Concordant (P), Discordant (Q), Tied (T_x,T_y), and Total Paired Observations: KeyDesign Example



the absolute value of gamma is the proportional reduction in error when prediction is done the second way. For example, you would get a 50 percent hit ratio using the coin. A PRE of .70 improves your hit ratio to 85 percent (.50 × 70) + (.50) = .85.

With a γ of –.70, 85 percent of the pairs are discordant and .15 percent are concordant.[16] There are almost six times as many discordant pairs as concordant pairs. In situations where the data call for a 2 × 2 table, the appropriate modification of gamma is Yule's *Q*.[17]

Kendall's **tau *b*** (τ_b) is a refinement of gamma that considers tied pairs. A tied pair occurs when subjects have the same value on the *X* variable, on the *Y* variable, or on both. For a given sample size, there are $n(n-1)/2$ pairs of observations.[18] After concordant pairs and discordant pairs are removed, the remainder are tied. Tau *b* does not have a PRE interpretation but does provide a range of –1.0 to +1.0 for square tables. Its compensation for ties uses the information found in Exhibit 18–24. It may be calculated as

$$\tau_b = \frac{P - Q}{\sqrt{\left(\frac{n(n-1)}{2} - T_x\right)\left(\frac{n(n-1)}{2} - T_y\right)}}$$

$$= \frac{172 - 984}{\sqrt{(2415 - 871)(2415 - 791)}} = -.5128$$

Kendall's **tau *c*** (τ_c) is another adjustment to the basic $P - Q$ relationship of gamma. This approach to ordinal association is suitable for tables of any size. Although we illustrate tau *c*, we would select tau *b* since the cross-classification table for the fitness data is square. The adjustment for table shape is seen in the formula

$$\tau_c = \frac{2m(P - Q)}{N^2(m - 1)} = \frac{2(3)(172 - 984)}{(70)^2(3 - 1)} - .4971$$

where *m* is the smaller number of rows or columns.

Somers's *d* rounds out our coverage of statistics employing the concept of concordant-discordant pairs. This statistic's utility comes from its ability to compensate for tied ranks and adjust for the direction of the dependent variable. Again, we refer to the preliminary calculations provided in Exhibit 18–24 to compute the symmetric and asymmetric *d*s. As before, the symmetric coefficient (equation 3) takes the row and column variables into account equally. The second and third calculations show fitness as the dependent and management level as the dependent, respectively.

$$d_{sym} = \frac{(P - Q)}{n(n-1) - T_x T_y/2} = \frac{-812}{1584} = -.5126 \quad \textbf{(3)}$$

$$d_{y-x} = \frac{(P - Q)}{\frac{n(n-1)}{2} - T_x} = \frac{-812}{2415 - 871} = -.5259 \quad \textbf{(4)}$$

$$d_{x-y} = \frac{(P - Q)}{\frac{n(n-1)}{2} - T_y} = \frac{-812}{2415 - 791} = -.5000 \quad \textbf{(5)}$$

The **Spearman's rho** (ρ) correlation is a popular ordinal measure. Along with Kendall's tau, it is among the most widely used of ordinal techniques. Rho correlates ranks between two ordered variables. Occasionally, researchers find continuous variables with too many abnormalities to correct. Then scores may be reduced to ranks and calculated with Spearman's rho.

As a special form of Pearson's product moment correlation, rho's strengths outweigh its weaknesses. When data are transformed by logs or squaring, rho remains unaffected. Second, outliers or extreme scores that were troublesome before ranking no longer pose a threat since the largest number in the distribution is equal to the sample size. Third, it is an easy statistic to compute. The major deficiency is its sensitivity to tied ranks. Too many ties distort the coefficient's size. However, there are rarely too many ties to justify the correction formulas available.

To illustrate the use of rho, consider a situation where Dean Merrill, a brokerage firm, is recruiting account executive trainees. Assume the field has been narrowed to 10 applicants for final evaluation. They arrive at the company headquarters, go through a battery of tests, and are interviewed by a panel of three executives. The test results are evaluated by an industrial psychologist who then ranks the 10 candidates. The executives produce a composite ranking based on the interviews. Your task is to decide how well these two sets of ranking agree. Exhibit 18–25 contains the data and preliminary calculations. Substituting into the equation, we get

$$r_s = 1 - \frac{6\Sigma d^2}{n^3 - n} = \frac{6(57)}{(10)^3 - 10} = .654$$

where n is the number of subjects being ranked.

The relationship between the panel's and the psychologist's ranking is moderately high, suggesting agreement between the two measures. The test of the null hypothesis that there is no relationship between the measures ($r_s = 0$) is rejected at the .05 level with $n - 2$ degrees of freedom.

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}} = \sqrt{\frac{8}{1-.4277}} = 2.45$$

EXHIBIT 18–25 Dean Merrill Data for Spearman's rho

Applicant	Rank by Panel x	Rank by Psychologist y	d	d^2
1	3.5	6	–2.5	6.25
2	10	5	5	25.00
3	6.5	8	–1.5	2.25
4	2	1.5	0.5	0.25
5	1	3	–2	4.00
6	9	7	2	4.00
7	3.5	1.5	2	4.00
8	6.5	9	–2.5	6.25
9	8	10	–2	4.00
10	5	4	1	1.00
				57.00

Note: Tied ranks were assigned the average (of ranks) as if no ties had occurred.

SUMMARY

1 Management questions frequently involve relationships between two or more variables. Correlation analysis may be applied to study such relationships. A correct correlational hypothesis states that the variables occur together in some specified manner *without* implying that one causes the other.

2 Parametric correlation requires two continuous variables measured on an interval or ratio scale. The product moment correlation coefficient represents an index of the magnitude of the relationship: Its sign governs the direction and its square explains the common variance. Bivariate correlation treats X and Y variables symmetrically and is intended for use with variables that are linearly related.

Scatterplots allow the researcher to visually inspect relationship data for appropriateness of the selected statistic. The direction, magnitude, and shape of a relationship are conveyed in a plot. The shape of linear relationships is characterized by a straight line, whereas nonlinear relationships are curvilinear or parabolic or have other curvature. The assumptions of linearity and bivariate normal distribution may be checked through plots and diagnostic tests.

A correlation matrix is a table used to display coefficients for more than two variables. Matrices form the basis for computation and understanding of the nature of relationships in multiple regression, discriminant analysis, factor analysis, and many multivariate techniques.

A correlation coefficient of any magnitude or sign, regardless of statistical significance, does not imply causation. Similarly, a coefficient is not remarkable simply because it is statistically significant. Practical significance should be considered in interpreting and reporting findings.

3 Regression analysis is used to further our insight into the relationship of Y with X. When we take the observed values of X to estimate or predict corresponding Y values, the process is called simple prediction. When more than one X variable is used, the outcome is a function of multiple predictors. Simple and multiple predictions are made with regression analysis.

A straight line is fundamentally the best way to model the relationship between two continuous variables. The method of least squares allows us to find a regression line, or line of best fit, that minimizes errors in drawing the line. It uses the criterion of minimizing the total squared errors of estimate. Point predictions made from well-fitted data are subject to error. Prediction and confidence bands may be used to find a range of probable values for Y based on the chosen predictor. The bands are shaped in such a way that predictors farther from the mean have larger band widths.

4 We test regression models for linearity and to discover whether the equation is effective in fitting the data. An important test in bivariate linear regression is whether the slope is equal to zero. In bivariate regression, t-tests and F tests of the regression produce the same result since t^2 is equal to F.

5 Often the assumptions or the required measurement level for parametric techniques cannot be met. Nonparametric measures of association offer alternatives. Nominal measures of association are used to assess the strength of relationships in cross-classification tables. They are often used in conjunction with chi-square or may be based on the proportional reduction in error (PRE) approach.

Phi ranges from 0 to +1.0 and attempts to correct chi-square proportionately to *N*. Phi is best employed with 2 × 2 tables. Cramer's *V* is a modification of phi for larger tables and has a range up to 1.0 for tables of any configuration. Lambda, a PRE statistic, is based on how well the frequencies of one nominal variable offer predictive evidence about the frequencies of another. Goodman and Kruskal's tau uses table marginals to reduce prediction errors.

Measures for ordinal data include gamma, Kendall's tau *b* and tau *c*, Somers's *d*, and Spearman's rho. All but Spearman's rank-order correlation are based on the concept of concordant and discordant pairs. None of these statistics require the assumption of a bivariate normal distribution, yet by incorporating order, most produce a range from −1 to +1.

KEY TERMS

EXAMPLES

Company	Scenario	Page
Containers Inc.*	Implementing a smoke-free workplace policy by evaluating the relationship between accidents and smoking.	592
Dean Merrill*	A brokerage firm uses an ordinal measure of association in recruiting account executive trainees.	600
European MBA programs	Specialty programs in business administration reflect emerging market niches.	582
Forbes 500 data	The relationship between cash flow and net profits in 10 companies.	575
KeyDesign*	Managerial employees at a large industrial design firm are evaluated for coronary risk.	596
McDonald's	Evaluating the effectiveness of a cashless payment system.	596
Mobil Oil Corp. (Speedpass)	Originator of the Speedpass system now being tested by select McDonald's restaurants in the Chicagoland area.	596
UCLA, U of PA, UC-Berkeley, Columbia, Harvard, Yale, and Northwestern universities	Business schools with wine clubs.	587

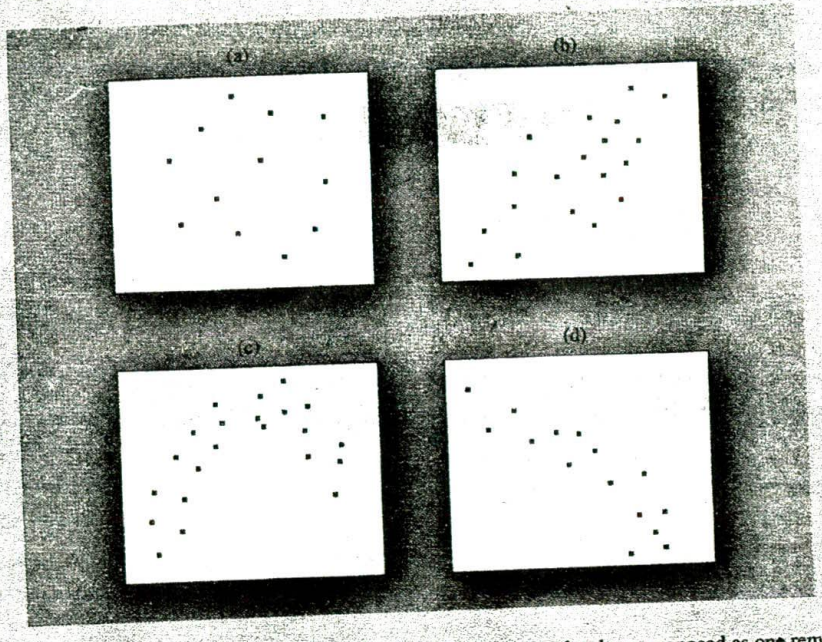
*Due to the confidential and proprietary nature of most research, the names of some companies have been changed.

DISCUSSION QUESTIONS

Terms in Review

1. Distinguish between the following:
   a. Regression coefficient and correlation coefficient.
   b. $r = 0$ and $\rho = 0$.
   c. The test of the true slope, the test of the intercept, and $r^2 = 0$.
   d. r^2 and r.
   e. A slope of 0.
   f. F and t^2.

2. Describe the relationship between the two variables in the four plots.



Making Research Decisions

3. A tax on the market value of stock and bond transactions has been proposed as one remedy for the budget deficit. The following data were collected on a sample of 60 registered voters by a polling organization.

	Education		
Opinion About Market Tax	H.S.	College Grad.	MBA
Favorable	15	5	0
Undecided	10	8	2
Unfavorable	0	2	18

   a. Compute gamma for the table.
   b. Compute tau b or tau c for the same data.

c. What accounts for the differences?

d. Decide which is more suitable for this data.

4. Using the table data in question 3, compute Somers's *d* symmetric and then use opinion as the dependent variable. Decide which approach is best for reporting the decision.

5. A research team conducted a study of voting preferences among 130 registered Democrats and 130 registered Republicans before an election on a specific tax proposal. They secured the following results:

	Favor	Against
Democrats	50	80
Republicans	90	40

Calculate an appropriate measure of association and decide how to present your results.

From Concept to Practice

6. Using this data,

X	Y
3	6
6	10
9	15
12	24
15	21
18	20

a. Create a scatterplot.

b. Find the least-squares line.

c. Plot the line on the diagram.

d. Predict: *Y* if *X* is 10.
Y if *X* is 17.

7. A home pregnancy test claims to be 97 percent accurate when consumers obtain a positive result. To what extent are the variables of "actual clinical condition" and "test readings" related?

a. Compute phi, Cramer's *V*, and the contingency coefficient for the table below. What can you say about the strength of the relationship between the two variables?

b. Compute lambda for this data. What does this statistic tell you?

*Actual Clinical Condition * Test Readings of In-Vitro Diagnostic Cross-Tabulaion*

Count		Test Readings of In-Vitro Diagnostic: Positive	Negative	Total
Actual clinical condition	Pregnant	451 accurate	36 innacurate	487
	Not pregnant	15 innacurate	183 accurate	198
Total		466	219	685

8. Fill in the missing blocks for the ANOVA summary table on net profits and market value used with regression analysis.

ANOVA Summary Table

	d.f.	Sum of Squares	Mean Square	*F*
Regression	1	11116995.47		
Error			116104.63	
Total	9	12045832.50		

a. What does the *F* tell you? ($\alpha = .05$).

b. What is the *t* value? Explain its meaning.

9. Using a computer program, produce a correlation matrix for the following data:

Forbes 500 Random Subsample (dollars in millions)

Assets	Sales	Market Value	Net Profit	Cash Flow	Number of Employees (thousands)
1034.00	1510.00	697.00	82.60	126.50	16.60
956.00	785.00	1271.00	89.00	191.20	5.00
1890.00	2533.00	1783.00	176.00	267.00	44.00
1133.00	532.00	752.00	82.30	137.10	2.10
11682.00	3790.00	4149.00	413.50	806.80	11.90
6080.00	635.00	291.00	18.10	35.20	3.70
31044.00	3296.00	2705.00	337.30	425.50	20.10
5878.00	3204.00	2100.00	145.80	380.00	10.80
1721.00	981.00	1573.00	172.60	326.60	1.90
2135.00	2268.00	2634.00	247.20	355.50	21.20

10. Secure Spearman rank-order correlations for the largest Pearson coefficient in the matrix from question 9. Explain the differences between the two findings.

11. Using the matrix data in question 9, select a pair of variables and run a simple regression. Then investigate the appropriateness of the model for the data using diagnostic tools for evaluating assumptions.

12. For the data below,

X	*Y*	*X*	*Y*
25	5	9	25
19	7	8	26
17	12	7	28
14	23	3	20
12	20		

a. Calculate the correlation between X and Y.

b. Interpret the sign of the correlation.

c. Interpret the square of the correlation.

d. Plot the least-squares line.

e. Test for a linear relationship.

(1) $\beta_1 = 0$.

(2) $r = 0$.

(3) An F test.

WWW Exercises Visit our website for Internet exercises related to this chapter at www.mhhe.com/business/cooper8

CASES

MASTERING TEACHER LEADERSHIP

PERFORMANCE EVALUATIONS

NCR: TEEING UP A NEW STRATEGIC DIRECTION

THE BRAZING OPERATION

OVERDUE BILLS

WORKLOAD RATINGS

All cases indicating a video icon are located on the Instructor's Videotape Supplement. All nonvideo cases are in the case section of the textbook. All cases indicating a CD icon offer a data set, which is located on the accompanying CD.

REFERENCE NOTES

1. Typically, we plot the x (independent) variable on the horizontal axis and the y (dependent) variable on the vertical axis. Although correlation does not distinguish between independent and dependent variables, the convention is useful for consistency in plotting and will be used later with regression.
2. F. J. Anscombe, "Graphs in Statistical Analysis," *American Statistician* 27 (1973), pp. 17–21. Cited in Samprit Chatterjee and Bertram Price, *Regression Analysis by Example* (New York: Wiley, 1977), pp. 7–9.
3. Amir D. Aczel, *Complete Business Statistics*, 2nd ed. (Homewood, IL: Irwin, 1993), p. 433.
4. The coefficient for net profits and cash flow in the example calculation used a subsample ($n = 10$) and was found to be .93. The matrix shows the coefficient as .95. The matrix calculation was based on the larger sample ($n = 100$).
5. This section is partially based on the concepts developed by Emanuel J. Mason and William J. Bramble, *Understanding and Conducting Research* (New York: McGraw-Hill, 1989), pp. 172–82; and elaborated in greater detail by Aczel, *Complete Business Statistics*, pp. 414–29.
6. Technically, estimation uses a concurrent criterion variable where prediction uses a future criterion. The statistical procedure is the same in either case.
7. Peter Passell, "Can Math Predict a Wine? An Economist Takes a Swipe at Some Noses," *International Herald Tribune*, March 5, 1990, p. 1; Jacques Neher, "Top Quality Bordeaux Cellar Is an Excellent Buy," *International Herald Tribune*, July 9, 1990, p. 8.
8. See Alan Agresti and Barbara Finlay, *Statistical Methods for the Social Sciences* (San Francisco: Dellen Publishing, 1986), pp. 248–49. Also see the discussion of basic regression models in John Neter, William Wasserman, and Michael H. Kutner, *Applied Linear Statistical Models* (Homewood, IL: Irwin, 1990), pp. 23–49.
9. We distinguish between the error terms $\varepsilon_i = Y_i - E[Y_i]$ and the residual $e_i = (Y_i - \hat{Y}_i)$. The first is based on the vertical deviation of Y_i from the true regression line. It is unknown and estimated. The second is the vertical deviation of Y_i from the fitted $\hat{Y}$ on the estimated line. See Neter et al., *Applied Linear Statistical Models*, p. 47.
10. For further information on software-generated regression diagnostics, see the most current release of software manuals for SPSS, MINITAB, BMDP, and SAS.
11. Aczel, *Complete Business Statistics*, p. 434.
12. This calculation is normally listed as the standard error of the slope (SE B) on computer printouts. For these data it is further defined as:

$$s(b_1) + \frac{8}{\sqrt{SS_x}} = \frac{538.559}{\sqrt{198.249}} = 38.249$$

where

s = The standard error of estimate (and the square root of the mean square error of the regression)

SS_x = The sum of squares for the X variable

13. Computer printouts use uppercase (R^2) because most procedures are written to accept multiple and bivariate regression.
14. The table output for this section has been modified from SPSS and is described in Norusis/SPSS, Inc., *SPSS Base System User's Guide*. For further discussion and examples of nonparametric measures of association, see S. Siegel and N. J. Castellan, Jr., *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed. (New York: McGraw-Hill, 1988).
15. Calculation of concordant and discordant pairs is adapted from Agresti and Finlay, *Statistical Methods for the Social Sciences*, pp. 221–23.
16. We know that the percentage of concordant plus the percentage of discordant pairs sums to 1.0. We also know their difference is –.70. The only numbers satisfying these two conditions are .85 and .15 (.85 + .15 = 1.0, .15 – .85 = –.70).
17. G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics* (New York: Hafner, 1950).
18. M. G. Kendall, *Rank Correlation Methods*, 4th ed. (London: Charles W. Griffin, 1970).

REFERENCES FOR SNAPSHOTS AND CAPTIONS

Business School Wine Clubs

E. Zelko, "Graduates of Wine," *Wine Spectator* 24, no. 15 (January 2000), pp. 88–90.

Specialized MBAs

M. Rowe, "Learning the Business of Wine, Sports, and Music," *International Herald Tribune*, (May 14, 2001), p. 18.

Speedpass

"McDonald's Accepts Speedpass for Fast Stomach Fill Up," press release (http://www.cardfram.com/html/news/090800_1.htm).

"McDonald's Expands Cashless Test," *PROMO XTRA!*, June 4, 2001.

"Speedpass Expands to More than 400 McDonald's Restaurants in Chicagoland Area," press release (http://biz.yahoo.com/bw/010531/0270.html).

CLASSIC AND CONTEMPORARY READINGS

Aczel, Amir D., and Jayauel Sounderpandian. *Complete Business Statistics*. 5th ed. Chicago: Irwin/McGraw-Hill, 2001. The chapter on simple regression/ correlation has impeccable exposition and examples and is highly recommended.

Agresti, Alan, and Barbara Finlay. *Statistical Methods for the Social Sciences*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 1997. Very clear coverage of nonparametric measures of association.

Chatterjee, Samprit, and Bertram Price. *Regression Analysis by Example*. 3rd ed. New York: Wiley, 1999. Updated version of widely used examples textbook.

Cohen, Jacob, and Patricia Cohen. *Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 1983. A classic reference work.

Neter, John, Michael H. Kutner, Christopher J. Nachtsheim, and William Wasserman. *Applied Linear Statistical Models*. 4th ed. Burr Ridge, IL: Irwin, 1996. Chapters 1 through 10 and 15 provide an excellent introduction to regression and correlation analysis.

Siegel, S., and N. J. Castellan, Jr. *Nonparametric Statistics for the Behavioral Sciences*. 2nd ed. New York: McGraw-Hill, 1988.