Business Research Methods, 12e

Online Learning Center Supplement

Multivariate Analysis: An Overview

>learningobjectives

After reading this chapter, you should understand . . .

1. How to classify and select multivariate techniques.
2. How multiple regression predicts a metric dependent variable from a set of metric independent variables.
3. How discriminant analysis classifies people or objects into categorical groups using several metric predictors.
4. How multivariate analysis of variance assesses the relationship between two or more metric dependent variables and independent classificatory variables.
5. How structural equation modeling explains causality among constructs that cannot be directly measured.
6. How conjoint analysis assists researchers to discover the most important attributes and levels of desirable features.
7. How principal components analysis extracts uncorrelated factors from an initial set of variables and how (exploratory) factor analysis reduces the number of variables to discover underlying constructs.
8. How cluster analysis techniques identify homogenous groups of objects or people using a set of variables to compare their attributes and/or characteristics.
9. How perceptions of products or services are revealed numerically and geometrically by multidimensional scaling.

“Wonder, connected with a principle of rational curiosity, is the source of all knowledge and discovery . . . but wonder which ends in wonder, and is satisfied with wonder, is the quality of an idiot.”

Samuel Horsley, scientist and fellow
Royal Society

>bringingresearchtolife

Parker drapes his arm across Sara's shoulder, before bending in close to breathe his greeting in her face. "Saw some of my favorite people and just had to stop by for a 'friendly hello.' "

Jason takes pity on Sara, drawing Parker's attention as Sara tries to shrug off his arm. "How's business, Henry?" Jason inquires, although he already knows Parker's firm lost a proposed project to them just that morning. He stands and extends his hand for a handclasp he really doesn't want, with a quick smile thrown Sara's way that says, "You owe me!"

Parker clasps Jason's extended hand and puts a lock on his right bicep as well. Now it is Sara's turn to commiserate the invading of Jason's personal space.

It was Parker's annoying practice, while holding you in his firm grip, to make amazingly improbable comparisons between people, groups, institutions, products, services, practices—anything and everything—by declaring the likes of "All things being equal, Mercury would seem to be a more congenial planet on which life might emerge than Earth." Meaning, if you allowed for its atmosphere being nonexistent, and its temperature being 1,380 degrees Fahrenheit, there was presumably something about its gravitational fields or length of day that fitted Parker's preferred cosmology. You cannot argue against that kind of pseudoscientific blather.

Now Parker is lecturing Jason about a project he is doing with the governing board of the public housing authority. "The best tenants are the Pantamarians," he declares. "All things being equal, they are the most law-abiding and hard-working tenants. These folks are from Pantamarie, all English-speakers from a little island in the Caribbean. Never heard of Pantamarie before I started this project, but, I tell you, they are the most law-abiding tenants . . ."

". . . all things being equal," echoes Jason ironically, as the very same words slip from Parker's mouth. Sara sees signs of Jason's increasing impatience, as he struggles to free himself from Parker's grasp.

"Do be more specific," urges Jason, yanking his arm from Parker's grasp none too gently. "Are you telling me that the Pantamarians have the lowest crime rate in the housing authority? You must have data—your project's funded by federal funds, right? So you must have data."

"Well," says Parker, evasively, "you have got to allow for these Pantamarians having very large families. And they did not get much schooling, back home."

"So what is not equal is their family size and education. What else is not equal?" Jason leans forward into Parker's space and stares icily into Parker's eyes.

Unbeknownst to Parker, he is saved from Jason's impending verbal attack by the arrival of the waiter carrying a loaded lunch tray.

"Well, I see lunch has arrived . . . nice to see you all . . . enjoy," smiles Parker as he turns and walks away.

"Parker wouldn't know how to prove his Pantamarian theory if we ran the numbers for him," shares Jason to the table at large. "You can be sure that the authority staff has been keeping really good records—family size, education, age—the Federal Housing and Urban Development people won't give Parker's firm a penny without it. But I'm equally sure he hasn't accessed those data.

"So, David, what would you do to prove or disprove Parker's theory?"

David, a doctoral student interning for the semester, pauses in lifting the fork to his lips. "I'd set crime rate as the dependent variable and country of origin as the independent variable and apply *analysis of covariance,* correcting for the effects of education, age, household size, whatever."

"Or maybe he could do a *factor analysis* that includes Caribbean country of origin, the population count for 2005, GDP per capita, teacher ratios, female life expectancy, births and deaths, the infant mortality rate per 1,000 of the population, radios and phones per 100 people, hospital beds, age, and family size. Then he'd know which variables are worth studying."

"Better yet," contributes Sara, joining into the spirit of the exercise Jason has started for his intern, "Parker could take the results of your factor analysis and run a *multiple regression* with crime rate as the dependent variable and the new factors that we output from the factor analysis as predictors."

"What about this," Jason contributes with a grin. "Parker could take his famous Pantamarians and the same data for their neighboring countrymen and see if he could correctly classify them with a *discriminant analysis.* Voilà! His Pantamarians could be proved to be the most law-abiding tenants," Jason pauses for effect. "Or not—all things being equal!"

Jason grins at Sara. "I completely forgot to congratulate him on landing the public authority contract and losing the more lucrative one—to us!"

After pausing for effect, Sara asks, "Now, David, what was that you were saying about your *multidimensional scaling* problem before Parker interrupted?"

> Introduction

In recent years, multivariate statistical tools have been applied with increasing frequency to research problems. This recognizes that many problems we encounter are more complex than the problems bivariate models can explain. Simultaneously, computer programs have taken advantage of the complex mathematics needed to manage multiple-variable relationships. Today, computers with fast processing speeds and versatile software bring these powerful techniques to researchers.

Throughout business, more and more problems are being addressed by considering multiple independent and/or multiple dependent variables. Sales managers base forecasts on various product history variables; researchers consider the complex set of buyer preferences and preferred product options; and analysts classify levels of risk based on a set of predictors.

One author defines **multivariate analysis** as "those statistical techniques which focus upon, and bring out in bold relief, the structure of simultaneous relationships among three or more phenomena."[1] Our overview of multivariate analysis seeks to illustrate the meaning of this definition while building on your understanding of bivariate statistics from the last few chapters. Several common multivariate techniques and examples will be discussed.

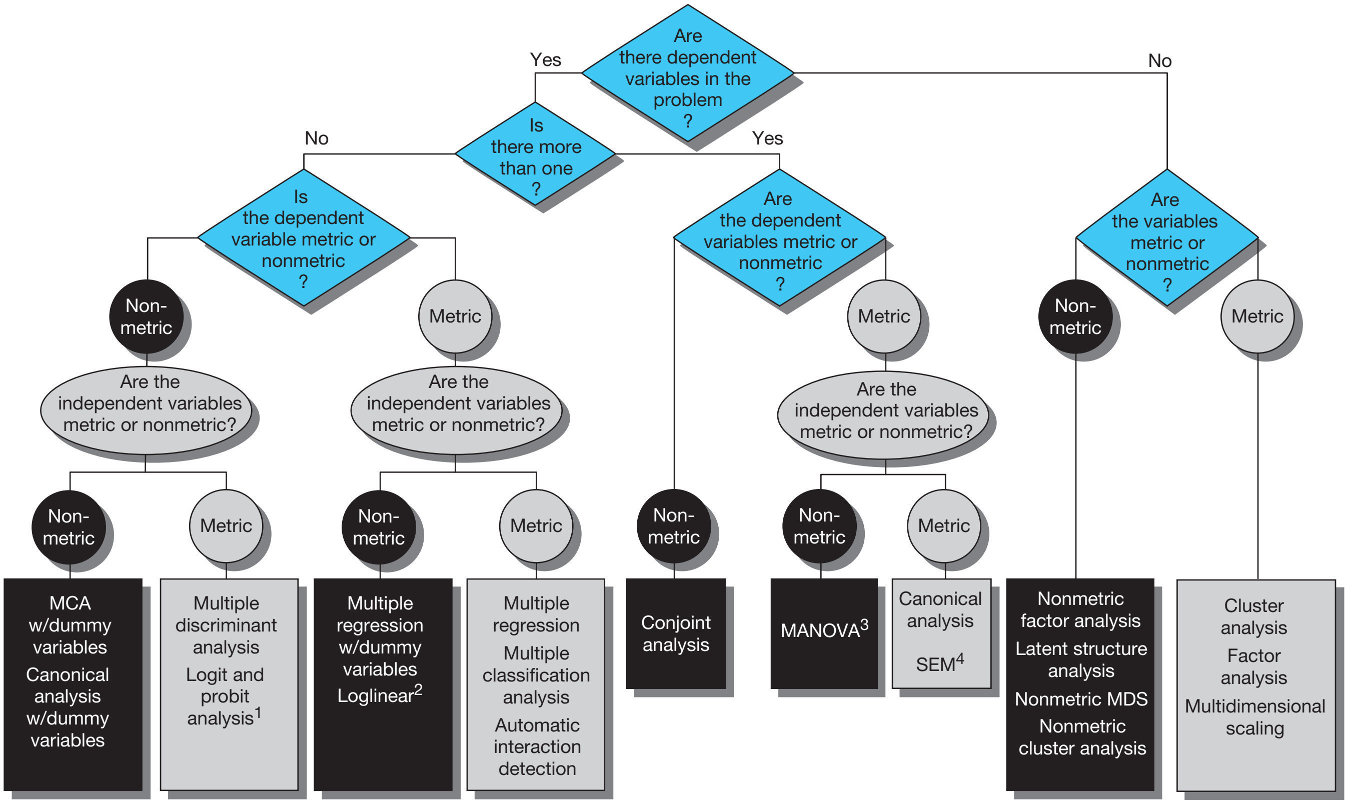
Because a complete treatment of this subject requires a thorough consideration of the mathematics, assumptions, and diagnostic tools appropriate for each technique, our coverage is necessarily limited. Readers desiring greater detail are referred to the suggested readings for this chapter.

> Selecting a Multivariate Technique

Multivariate techniques may be classified as **dependency** and **interdependency techniques.** Selecting an appropriate technique starts with an understanding of this distinction. If criterion and predictor variables exist in the research question, then we will have an assumption of dependence. Multiple regression, multivariate analysis of variance (MANOVA), and discriminant analysis are techniques in which criterion or dependent variables and predictor or independent variables are present. Alternatively, if the variables are interrelated without designating some as dependent and others independent, then interdependence of the variables is assumed. Factor analysis, cluster analysis, and multidimensional scaling are examples of interdependency techniques.

Exhibit MV-1 provides a diagram to guide in the selection of techniques. Let's take an example to show how you might make a decision. Every other year since 1978, the Roper organization has tracked public opinion toward business by providing a list of items that are said to be the responsibility of

>Exhibit MV-1 Selecting from the Most Common Multivariate Techniques



[1]The independent variable is metric only in the sense that a transformed proportion is used.
[2]The independent variable is metric only when we consider that the number of cases in the cross-tabulation cell is used to calculate the logs.
[3]Factors may be considered nonmetric independent variables in that they organize the data into groups. We do not classify MANOVA and other multivariate analysis of variance models.
[4]SEM refers to structural equation modeling for latent variables. It is a family of models appropriate for confirmatory factor analysis, path analysis, time series analysis, recursive and nonrecursive models, and covariance structure models. Because it may handle dependence and interdependence, metric and nonmetric, it is arbitrarily placed in this diagram.

Source: Partially adaxpted from T. C. Kinnear and J. R. Taylor, "Multivariate Methods in Marketing: A Further Attempt at Classification," *Journal of Marketing,* October 1971, p. 57; and J. F. Hair Jr., Rolph E. Anderson, Ronald L. Tatham, and Bernie J. Grablowsky, *Multivariate Data Analysis* (Tulsa, OK: Petroleum Publishing Co., 1979), pp. 10–14.

business. The respondents are asked whether business fulfills these responsibilities "fully, fairly well, not too well, or not at all well." The following issues make up the list:[2]

- Developing new products and services.
- Producing good-quality products and services.
- Making products that are safe to use.
- Hiring minorities.
- Providing jobs for people.
- Being good citizens of the communities in which they operate.
- Paying good salaries and benefits to employees.
- Charging reasonable prices for goods and services.
- Keeping profits at reasonable levels.
- Advertising honestly.

- Paying their fair share.
- Cleaning up their own air and water pollution.

You have access to data on these items and wish to know if they could be reduced to a smaller set of variables that would account for most of the variation among respondents. In response to the first question in Exhibit MV-1, you correctly determine there are no dependent variables in the data set. You then check to see if the variables are **metric** or **nonmetric measures.** In the exhibit, *metric* refers to ratio and interval measurements, and *nonmetric* refers to data that are nominal and ordinal. Based on the measurement scale, which appears to have equal intervals, and preliminary findings that show a linear relationship among several variables, you decide the data are metric. This decision leads you to three options: multidimensional scaling, cluster analysis, or factor analysis. *Multidimensional scaling* develops a perceptual map of the locations of some objects relative to others. This map specifies how the objects differ. *Cluster analysis* identifies homogeneous subgroups or clusters of individuals or objects based on a set of characteristics. *Factor analysis* looks for patterns among the variables to discover if an underlying combination of the original variables (a factor) can summarize the original set. Based on your research objective, you select factor analysis.

Suppose you are interested in predicting family food expenditures from family income, family size, and whether the family's location is rural or urban. Returning to Exhibit MV-1, you conclude there is a single dependent variable, family food expenditures. You decide this variable is metric because dollars are measured on a ratio scale. The independent variables, income and family size, also meet the criteria for metric data. However, you are not sure about the location variable because it appears to be a dichotomous nominal variable. According to the exhibit, your choices are automatic interaction detection (AID), multiple classification analysis (MCA), and multiple regression. You recall from Chapter 16 that AID was designed to locate the most important predictors in a set of numerous independent variables and create a treelike answer. MCA handles weak predictors (including nominal variables), correlated predictors, and nonlinear relationships. Multiple regression is the extension of bivariate regression. You believe that your data exceed the assumptions for the first two techniques and that by treating the nominal variable's values as 0 or 1, you could use it as an independent variable in a multiple regression model. You prefer this to losing information from the other two variables—a certainty if you reduce them to nonmetric data.

In the next two sections, we will extend this discussion as we illustrate dependency and interdependency techniques.[3]

> Dependency Techniques

Multiple Regression

Multiple regression is used as a descriptive tool in three types of situations. First, it is often used to develop a self-weighting estimating equation by which to predict values for a criterion variable (DV) from the values for several predictor variables (IVs). Thus, we might try to predict company sales on the basis of new housing starts, new marriage rates, annual disposable income, and a time factor. Another prediction study might be one in which we estimate a student's academic performance in college from the variables of rank in high school class, SAT verbal scores, SAT quantitative scores, and a rating scale reflecting impressions from an interview.

Second, a descriptive application of multiple regression calls for controlling for confounding variables to better evaluate the contribution of other variables. For example, one might wish to control the brand of a product and the store in which it is bought to study the effects of price as an indicator of product quality.[4] A third use of multiple regression is to test and explain causal theories. In this approach, often referred to as **path analysis,** regression is used to describe an entire structure of linkages that have been advanced from a causal theory.[5] In addition to being a descriptive tool, multiple regression is also used as an inference tool to test hypotheses and to estimate population values.

Method

Multiple regression is an extension of the bivariate linear regression presented in Chapter 18. The terms defined in that chapter will not be repeated here. Although **dummy variables** (nominal variables coded 0, 1) may be used, all other variables must be interval or ratio. The generalized equation is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

where

β_0 = a constant, the value of Y when all X values are zero

β_i = the slope of the regression surface (The β represents the regression coefficient associated with each X_i.)

ε = an error term, normally distributed about a mean of 0 (For purposes of computation, the ε is assumed to be 0.)

The regression coefficients are stated either in raw score units (the actual X values) or as **standardized coefficients** (X values restated in terms of their standard scores). In either case, the value of the regression coefficient states the amount that Y varies with each unit change of the associated X variable when the effects of all other X variables are being held constant. When the regression coefficients are standardized, they are called **beta weights** (β), and their values indicate the relative importance of the associated X values, particularly when the predictors are unrelated. For example, in an equation where $\beta_1 = .60$ and $\beta_2 = .20$, one concludes that X_1 has three times the influence on Y as does X_2.

Example

In a Snapshot later in this chapter, we describe an e-business that uses multivariate approaches to understand its target market in the global “hybrid-mail” business. SuperLetter’s basic service enables users to create a document on any PC and send it in a secure, encrypted mode over the Internet to a distant international terminal near the addressee, where it will be printed, processed, and delivered. Spread like a “fishnet” over the world’s major commercial markets, the network connects corresponding parties, linking the world’s “wired” with its “nonwired.” The British Armed Forces and several U.S. military organizations have used it to speed correspondence between families and service members in Afghanistan and Iraq.

We use multiple regression in this example to evaluate the *key drivers* of customer usage for hybrid mail. Among the available independent or predictor variables, we expect some to better explain or predict the dependent or criterion variable than others (thus they are *key* to our understanding). The independent variables are customer perceptions of (1) cost/speed valuation, (2) security (limits on changing, editing, or forwarding a document and document privacy), (3) reliability, (4) receiver technology (hard copy for receivers with no e-mail or fax access), and (5) impact/emotional value (reducing e-mail spam clutter and official/important appearance). We have chosen the first three variables, all measured on 5-point scales, for this equation:

Y = customer usage

X_1 = cost/speed valuation

X_2 = security

X_3 = reliability

SPSS computed the model and the regression coefficients. Most statistical packages provide various methods for selecting variables for the equation. The equation can be built with all variables or specific combinations, or you can select a method that sequentially adds or removes variables (forward selection, backward elimination, and stepwise selection). **Forward selection** starts with the constant and adds variables one at a time that result in the largest R^2 increase. **Backward elimination** begins with a model containing all independent variables and removes the variable that changes R^2 the least. **Stepwise selection,** the most popular method, combines forward and backward sequential approaches.

>Exhibit MV-2 Multiple Regression Analysis of Hybrid-Mail Customer Usage, Cost/Speed Valuation, Security, and Reliability

Model Summary

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate	Change Statistics: R^2 Change	F Change	d.f.1	d.f.2	Sig. F Change
1	.879	.772	.771	.6589	.772	612.696	1	181	.000
2	.925	.855	.854	.5263	.083	103.677	2	180	.000
3	.935	.873	.871	.4937	.018	25.597	3	179	.000

1 Predictors: (constant), cost/speed.
2 Predictors: (constant), cost/speed, security.
3 Predictors: (constant), cost/speed, security, reliability.

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics
Model		B	Std. Error	Beta			VIF
1	(Constant)	.579	.151		3.834	.000	
	Cost/speed	.857	.035	.879	24.753	.000	1.000
2	(Constant)	9.501E-02	.130		.733	.464	
	Cost/speed	.537	.042	.551	12.842	.000	2.289
	Security	.428	.042	.437	10.182	.000	2.289
3	(Constant)	–9.326E-02	.127		–.734	.464	
	Cost/speed	.448	.043	.460	10.428	.000	2.748
	Security	.315	.045	.321	6.948	.000	3.025
	Reliability	.254	.050	.236	5.059	.000	3.067

Dependent variable: customer usage.

The independent variable that contributes the most to explaining the dependent variable is added first. Subsequent variables are included based on their incremental contribution over the first variable and on whether they meet the criterion for entering the equation (e.g., a significance level of .01). Unlike forward selection and backward elimination, stepwise selection allows variables to be added or deleted at each subsequent step in the method. Variables may be removed if they meet the removal criterion, which is a larger significance level than that for entry.

The standard elements of a stepwise output are shown in Exhibit MV-2. In the upper portion of the exhibit there are three models. In model 1, cost/speed is the first variable to enter the equation. This model consists of the constant and the variable cost/speed. Model 2 adds the security variable to cost/speed. Model 3 consists of all three independent variables. In the summary statistics for model 1, you see that cost/speed explains 77 percent of customer usage (see the "R^2" column). This is increased by 8 percent in model 2 when security is added (see "R^2 Change" column). When reliability is added in model 3, accounting for only 2 percent, 87 percent of customer usage is explained.

The other reported statistics have the following interpretations.

1. Adjusted R^2 for model 3 = .871. R^2 is adjusted to reflect the model's goodness of fit for the population. The net effect of this adjustment is to reduce the R^2 from .873 to .871, thereby making it comparable to other R^2's from equations with a different number of independent variables.
2. Standard error of model 3 = .4937. This is the standard deviation of actual values of Y about the estimated Y values.

3. Analysis of variance measures whether or not the equation represents a set of regression coefficients that, in total, are statistically significant from zero. The critical value for F is found in Appendix D (Exhibit D-8), with degrees of freedom for the numerator equaling k, the number of independent variables, and for the denominator, $n - k - 1$, where n for model 3 is 183 observations. Thus, d.f. = (3, 179). The equation is statistically significant at less than the .05 level of significance (see the column labeled "Sig. F Change").
4. Regression coefficients for all three models are shown in the lower table of Exhibit MV-2. The column headed "B" shows the unstandardized regression coefficients for the equation. The equation may now be constructed as

$$Y = -.093 + .448X_1 + .315X_2 + .254X_3$$

5. The column headed "Beta" gives the regression coefficients expressed in standardized form. When these are used, the regression Y intercept is zero. Standardized coefficients are useful when the variables are measured on different scales. The beta coefficients also show the relative contribution of the three independent variables to the explanatory power of this equation. The cost/speed valuation variable explains more than either of the other two variables.
6. Standard error is a measure of the sampling variability of each regression coefficient.
7. The column headed "t" measures the statistical significance of each of the regression coefficients.

Again compare these to the table of t values in Appendix D, Exhibit D-2, using degrees of freedom for one independent variable. All three regression coefficients are judged to be significantly different from zero. Therefore, the regression equation shows the relationship between the dependent variable, customer usage of hybrid mail, and three independent variables: cost/speed, security, and reliability. The regression coefficients are both individually and jointly statistically significant. The independent variable cost/speed influences customer usage the most, followed by security and then reliability.

Collinearity, where two independent variables are highly correlated—or **multicollinearity,** where more than two independent variables are highly correlated—can have damaging effects on multiple regression. When this condition exists, the estimated regression coefficients can fluctuate widely from sample to sample, making it risky to interpret the coefficients as an indicator of the relative importance of predictor variables. Just how high can acceptable correlations be between independent variables? There is no definitive answer, but, as a rule of thumb, correlations at a .80 or greater level should be addressed. Because high intercorrelations between predictor variables suggest that they are measuring the same construct, the presence of multicollinearity can be dealt with in one of two ways:

1. Retain the variable that best captures the concept/construct you want to measure and delete the other.
2. Create a new variable that is a composite of the highly intercorrelated variables and use this new variable in place of its components.

However, making a decision to delete or alter variables based on the findings contained within the correlation matrix alone is not always advisable. In the example just presented, Exhibit MV-2 contains a column labeled "Collinearity Statistics" that shows a *variable inflation factor (VIF)* index. This is a measure of the effect of the other independent variables on a regression coefficient as a result of these correlations. Large values, usually 10.0 or more, suggest collinearity or multicollinearity. For the three predictors in the hybrid-mail example, multicollinearity is not a problem. However, there may be instances when you determine that the predictor variables are, in fact, conceptually distinct (precluding deletion) yet are nevertheless highly correlated (such as income and education level). In this situation, you must acknowledge the possible impact of multicollinearity on the unique effects of each predictor variable.

The last step in the multiple regression technique is to evaluate (validate) how well the regression equation predicts beyond the sample used originally to calculate it. A practical solution is to set aside a portion of the data (from one-fourth to one-third) called a **holdout sample.** One uses the equation with the holdout data to calculate a new R^2 and compare its similarity to the original R^2 to determine if the results are generalizable to the population.

Discriminant Analysis

Researchers often wish to classify people or objects into two or more distinct and well-defined groups. One might need to classify persons as either buyers or nonbuyers, good or bad credit risks, or to classify superior, average, or poor products in some market. The objective of discriminant analysis is to establish a classification method, based on a set of attributes, in order to correctly predict the group membership of these subjects. With this objective, it is easy to understand why discriminant analysis is frequently used in market segmentation research.

Method

As a dependency technique, **discriminant analysis** joins a nominally scaled criterion or dependent variable with one or more independent variables that are interval- or ratio-scaled. Once the discriminant equation is determined, it can be used to predict the classification of a new observation. This is done by calculating a linear function of the form

$$D_i = d_0 + d_1X_1 + d_2X_2 + \cdots + d_pX_p$$

where

D_i is the score on discriminant function i.

The d_i's are weighting coefficients; d_0 is a constant.

The X's are the values of the discriminating variables used in the analysis.

A single discriminant equation is required if the categorization calls for two groups. If three groups are involved in the classification, it requires two discriminant equations. If more categories are called for in the dependent variable, one needs $N - 1$ discriminant functions. Of note, each of the discriminant groups are both collectively exhaustive and mutually exclusive, in that each entity belongs to a group and to only one group.

While the most common use for discriminant analysis is to classify persons or objects into various groups, it can also be used to analyze known groups to determine the relative influence of specific factors for deciding into which group various cases fall. Assume we have MindWriter service ratings that enable us to classify postpurchase service as successful or unsuccessful on performance. We might also be able to secure test results on three measures: motivation for working with customers (X_1), technical expertise (X_2), and accessibility to repair status information (X_3). Suppose the discriminant equation is

$$D = .06X_1 + .45X_2 + .30X_3$$

Since discriminant analysis uses standardized values for the discriminant variables, we conclude from the coefficients that motivation for working with customers is less important than the other two in classifying postpurchase service.[6]

Example

An illustration of the method takes us back to the problem in the last chapter where KDL, a media firm, is hiring MBAs for its account executives program. Over the years the firm has had indifferent success with the selection process. You are asked to develop a procedure to improve the firm's current selection process. It appears that discriminant analysis is the most appropriate technique to fulfill this task. You begin by gathering data on 30 MBAs who have been hired by KDL in recent years. Fifteen of these new hires have become successful employees, while the other 15 were unsuccessful. The files provide the following information that can be used to conduct the analysis:

X_1 = years of prior work experience

X_2 = GPA in graduate program

X_3 = employment test scores

>Exhibit MV-3 Discriminant Analysis Classification Results at KDL Media

A.

Actual Group		Number of Cases	Predicted Success	
			0	1
Unsuccessful	0	15	13	2
			86.70%	13.30%
Successful	1	15	3	12
			20.00%	80.00

Note: Percent of "grouped" cases correctly classified: 83.33%.

B.

	Unstandardized	Standardized
X_1	.36084	.65927
X_2	2.61192	.57958
X_3	.53028	.97505
Constant	12.89685	

Discriminant analysis determines how well these three independent variables will correctly classify those who are judged successful from those judged unsuccessful. The classification results are shown in Exhibit MV-3. This indicates that 25 of the 30 ($30 - 3 - 2 = 25$) cases have been correctly classified using these three variables. In interpreting the discriminant function, it is also important to examine the misclassified cases for additional attributes or relationships that may improve the accuracy of the equation.

The standardized and unstandardized discriminant function coefficients are shown in part B of Exhibit MV-3. These results indicate that X_3 (the employment test) has the greatest discriminating power of the three attributes or independent variables used to assess the differences between these groups. Of the several significance tests that may be computed, Wilk's lambda has a chi-square transformation for testing the significance of the discriminant function. If computed for this example, it indicates that the equation is statistically significant at the $\alpha = .0004$ level. Using the discriminant equation,

$$D = .659X_1 + .580X_2 + .975X_3$$

you can now predict with a greater level of certainty whether future candidates are likely to be successful account executives.

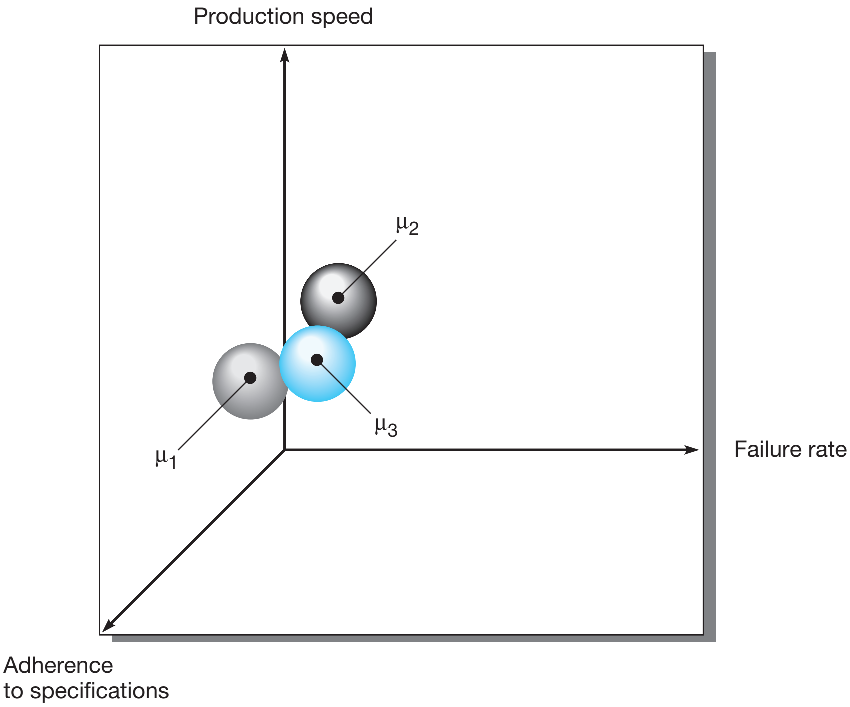
MANOVA

Multivariate analysis of variance, or **MANOVA,** is a commonly used multivariate technique. MANOVA assesses the relationship between two or more continuous dependent variables and categorical variables or factors. In business research, MANOVA can be used to test differences among samples of employees, customers, manufactured products, production parts, and so forth.

Method

MANOVA is similar to the univariate ANOVA described earlier, with the added ability to handle several dependent variables. If ANOVA is applied consecutively to a set of interrelated dependent variables, erroneous conclusions may result (e.g., Type I error). MANOVA can correct this by simultaneously testing all the variables and their interrelationships. MANOVA uses special matrices

>Exhibit MV-4 MANOVA Techniques Show These Three Centroids to Be Unequal in the CalAudio Study



[sums-of-squares and cross-products (SSCP) matrices] to test for differences among groups. The variance between groups is determined by partitioning the total SSCP matrix and testing for significance. The *F* ratio, generalized to a ratio of the within-group variance and total-group variance matrices, tests for equality among groups. MANOVA examines similarities and differences among the multivariate mean scores of several populations. The null hypothesis for MANOVA is that all of the **centroids** (multivariate means) are equal, H_0: $\mu_1 = \mu_2 = \mu_3 = \cdots \mu_n$. The alternative hypothesis is that the vectors of centroids are unequal, H_A: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \cdots \mu_n$. Exhibit MV-4 shows graphically three populations whose centroids are unequal, allowing the researcher to reject the null hypothesis. When the null hypothesis is rejected, additional tests are done to understand the results in detail. Several techniques may be considered in proceeding from a significant effect in MANOVA:

1. Univariate *F* tests can be run on the dependent variables (multiple univariate ANOVAs).
2. Simultaneous confidence intervals can be produced for each variable.
3. Stepdown analysis, like stepwise regression, can be run by computing *F* values successively. Each value is computed after the effects of the previous dependent variable are eliminated.
4. Multiple discriminant analysis can be used on the SSCP matrices. This aids in the discovery of which variables contribute to the MANOVA's significance.[7]

Before using MANOVA to test for significant differences, you must first determine that MANOVA is appropriate, that is, that the assumptions for its use are met.

Example

To illustrate how these assumptions are assessed and how MANOVA is performed, let's look at CalAudio, a firm that manufactures MP3 players. The manager is concerned about brand loyalty and fears that the quality of the manufactured players may be affecting customers' repurchase decisions. The closest competitor's product appears to have fewer repair issues and higher satisfaction ratings. Two measures are used to assess quality in this example: adherence to product specifications and time before failure. Measured on a 0-to-100 scale, with 100 meeting all product specifications, the specification variable is averaging approximately 90. The mean time before failure is calculated in weeks; it is approximately 159 weeks, or three years.

Management asks the industrial engineering department to devise a modified manufacturing procedure that will improve the quality measures but not change the production rate significantly. A new method is designed that includes more efficient parts handling and "burn-in" time, when MP3 players are powered up and run at high temperatures.

Engineering takes a sample of 15 MP3 players made with the old manufacturing method and 15 made with the new method. The players are measured for their adherence to product specifications and are stress-tested to determine their time before failure. The stress test uses accelerated running conditions and adverse environmental conditions to simulate years of use in a short time.

Exhibit MV-5 shows the mean and standard deviation of the dependent variables (failure, specifications, and manufacturing speed) for each level of method.[8] Method 1 represents the current manufacturing process, and method 2 is the new process. The new method extended the time before failure to 181 weeks, compared to 159 weeks for the existing method. The adherence to specifications is also improved, up to 95 from 90. But the manufacturing speed is slower by approximately 30 minutes (.473 hour).

We have used diagnostics to check the assumptions of MANOVA except for equality of variance. Both levels of the manufacturing method variable produce a matrix, and the equality of these two matrices must be determined (H_0: variances are equal). Exhibit MV-6 contains homogeneity-of-variance tests for separate dependent variables and a multivariate test. The former are known as *univariate tests.* The multivariate test is a comparable version that tests the variables simultaneously to determine whether MANOVA should proceed.

The significance levels of Cochran's *C* and Bartlett-Box *F* do not allow us to reject any of the tests for the dependent variables considered separately. This means the two methods have equal variances in each dependent variable. This fulfills the univariate assumptions for homogeneity of variance. We then consider the variances and covariances simultaneously with Box's *M,* also found in Exhibit MV-6. Again, we are unable to reject the homogeneity-of-variance assumption regarding the matrices. This satisfies the multivariate assumptions.

>Exhibit MV-5 MANOVA Cell Means and Standard Deviations in CalAudio Study

VARIABLE	FACTOR	LEVEL	MEAN	STD. DEV.
FAILURE				
	METHOD	1	158.867	4.998
	METHOD	2	181.067	5.994
	For entire sample		169.967	12.524
SPECIFICATIONS				
	METHOD	1	89.800	2.077
	METHOD	2	94.800	2.178
	For entire sample		92.300	3.292
SPEED				
	METHOD	1	2.126	.061
	METHOD	2	2.599	.068
	For entire sample		2.362	.249

>Exhibit MV-6 MANOVA Homogeneity-of-Variance Tests in the CalAudio Study

VARIABLE	TEST	RESULTS
FAILURE		
	Cochran's *C* (14,2) =	.58954, *P* = .506 (approx.)
	Bartlett-Box *F* (1,2352) =	.44347, *P* = .506
SPECIFICATIONS		
	Cochran's *C* (14,2) =	.52366, *P* = .862 (approx.)
	Bartlett-Box *F* (1,2352) =	.03029, *P* = .862
SPEED		
	Cochran's *C* (14,2) =	.55526, *P* = .684 (approx.)
	Bartlett-Box *F* (1,2352) =	.16608, *P* = .684

Multivariate Test for Homogeneity of Dispersion Matrices

Box's *M* =	6.07877
F with (6,5680) DF =	.89446, *P* = .498 (approx.)
Chi-Square with 6 DF =	5.37320, *P* = .497 (approx.)

>Exhibit MV-7 Bartlett's Test of Sphericity in the CalAudio Study

Statistics for WITHIN CELLS correlations

Log (Determinant) =	-3.92663
Bartlett's test of sphericity =	102.74687 with 3 D.F.
Significance =	.000
F(max) criterion =	7354.80161 with (3,28) D.F.

>Exhibit MV-8 Multivariate Tests of Significance in the CalAudio Study

Multivariate Tests of Significance (*S* = 1, *M* = 1/2, *N* = 12)

Test Name	Value	Exact *F*	Hypoth. DF	Error DF	Sig. of *F*
Hotelling	51.33492	444.90268	3.00	26.00	.000
Pillai	.98089	444.90268	3.00	26.00	.000
Wilks	.01911	444.90268	3.00	26.00	.000

Note: F statistics are exact.

When MANOVA is applied properly, the dependent variables are correlated. If the dependent variables are unrelated, there would be no necessity for a multivariate test, and we could use separate *F* tests for failure, specifications, and speed, much like the ANOVAs in Chapter 18. Bartlett's test of sphericity helps us decide if we should continue analyzing MANOVA results or return to separate univariate tests. In Exhibit MV-7, we will look for a determinant value that is close to 0. This implies that one or more dependent variables are a linear function of another. The determinant has a chi-square transformation that simplifies testing for statistical significance. Since the observed significance is below that set for the model ($\alpha = .05$), we are able to reject the null hypothesis and conclude there are dependencies among the failure, specifications, and speed variables.

We now move to the test of equality of means that considers the three dependent variables for the two levels of manufacturing method. This test is analogous to a *t*-test or an *F* test for multivariate data. The sums-of-squares and cross-products matrices are used. Exhibit MV-8 shows three tests, including the Hotelling T^2. All the tests provided are compared to the *F* distribution for interpretation. Since the observed significance level is less than $\alpha = .05$ for the T^2 test, we reject the null hypothesis that

>Exhibit MV-9 Univariate Tests of Significance in the CalAudio Study

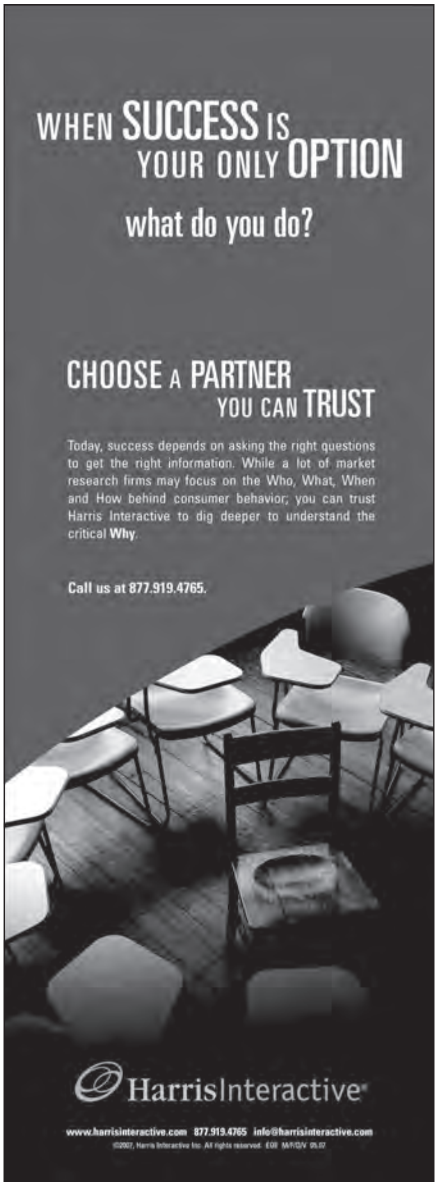
Univariate *F* Tests with (1,28) D.F.

Variable	Hypoth. *SS*	Error *SS*	Hypoth. *MS*	Error *MS*	*F*	Sig. of *F*
FAILURE	3696.30000	852.66667	3696.30000	30.45238	121.37967	.000
SPECS	187.50000	126.80000	187.50000	4.52857	41.40379	.000
SPEED	1.67560	.11593	1.67560	.00414	404.68856	.000

Note: F statistics are exact.

said methods 1 and 2 provide equal results with respect to failure, specifications, and speed. Similar results are obtained from the Pillai trace and Wilks's statistic.

Finally, to detect where the differences lie, we can examine the results of univariate *F* tests in Exhibit MV-9. Since there are only two methods, the *F* is equivalent to t^2 for a two-sample *t*-test. The significance levels for these tests do not reflect that several comparisons are being made, and we should use them principally for diagnostic purposes. This is similar to problems that require the use of multiple comparison tests in univariate analysis of variance. Note, however, that there are statistically significant differences in all three dependent variables resulting from the new manufacturing method. Techniques for further analysis of MANOVA results were listed at the beginning of this section.



Harris Interactive is one of the world's largest research firms. It is known for its sophisticated data analysis, as well as for its newsworthy Harris Poll. This ad stresses the importance of insight formation in solving strategic questions. **www.harrisinteractive.com**

Structural Equation Modeling[9]

Since the late 1980s, researchers have relied increasingly on structural equation modeling to test hypotheses about the dimensionality of, and relationships among, latent and observed variables. **Structural equation modeling (SEM)** implies a structure for the covariances between observed variables, and accordingly it is sometimes called *covariance structure modeling.* More commonly, researchers refer to structural equation models as LISREL (linear structural relations) models—the name of the first and most widely cited SEM computer program.

SEM is a powerful alternative to other multivariate techniques, which are limited to representing only a single relationship between the dependent and independent variables. The major advantages of SEM are (1) that multiple and interrelated dependence relationships can be estimated simultaneously and (2) that it can represent unobserved concepts, or *latent variables,* in these relationships and account for measurement error in the estimation process. While the details of SEM are quite complex, well beyond the scope of this text, this section provides a broad conceptual introduction.

Method

Researchers using SEM must follow five basic steps:

1. *Model specification.* The first step in SEM is the *specification,* or formal statement, of the model's *parameters.* These parameters, constants that describe the relations between variables, are specified as either *fixed* or *free.* Fixed parameters have values set by the researcher, and are not estimated from the data. For example, if there is no hypothesized relationship between variables, the parameter would be fixed at zero. When there is a hypothesized, but unknown, relation between the variables, the parameters are set free to be estimated from the data. Researchers must be careful to consider all the important predictive variables to avoid **specification error,** a bias that overestimates the importance of the variables included in the model.

2. *Estimation.* After the model has been specified, the researcher must obtain estimates of the free parameters from the observed data. This is often accomplished using an *iterative method,* such as *maximum likelihood estimation (MLE).*

3. *Evaluation of fit.* Following convergence, the researcher must evaluate the goodness-of-fit criteria. *Goodness-of-fit tests* are used to determine whether the model should or should not be rejected. If the model is not rejected, the researcher will continue the analysis and interpret the path coefficients in the model. Most, if not all, SEM computer software programs include several different goodness-of-fit measures, each of which can be categorized as one of three types of measures.

4. *Respecification of the model.* Model respecification usually follows the estimation of a model with indications of poor fit. Sometimes, the model is compared with competing or *nested* models to find the best fit among a set of models, and then the original model is respecified to produce a better fit. Respecifying the model requires that the researcher fix parameters that were formerly free or free parameters that were formerly fixed.

5. *Interpretation and communication.* SEM hypotheses and results are most commonly presented in the form of **path diagrams,** which are graphic illustrations of the measurement and structural models. The main features of path diagrams are ellipses, rectangles, and two types of arrows. The ellipses represent latent variables. Rectangles represent observed variables, which can be indicators of latent variables in the measurement model or of independent variables in the structural model. Straight arrows are pointed at one end and indicate the direction of prediction from independent to dependent variables or from indicators to latent variables. Curved arrows are pointed at both ends and indicate correlations between variables.

In a research report, the path diagrams should illustrate the model originally specified and estimated by the researcher; the portion of the model for which parameter estimates were significant; and a model that resulted from one or more modifications and reestimations of the original model. The researcher should also take care to include the method of estimation, the fit criteria selected, and the parameter estimates.

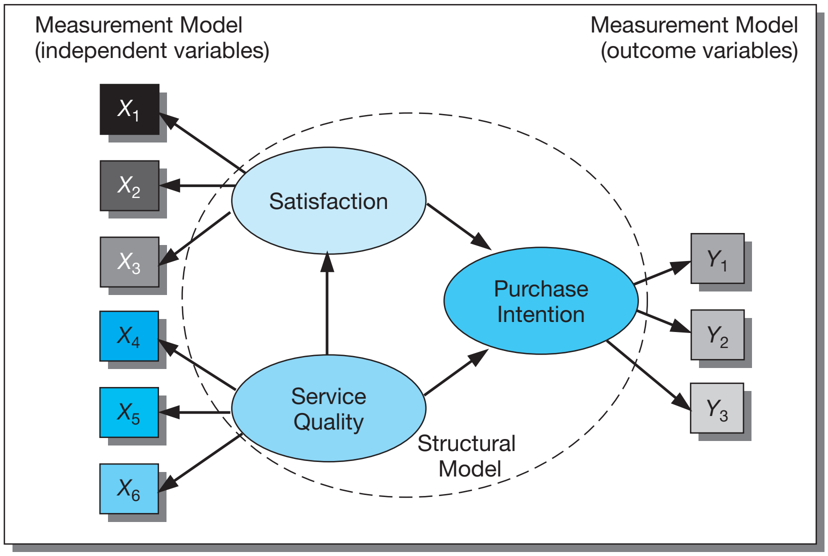
Example

A research consultant, hired by MindWriter, investigated the relationship between customer satisfaction and service quality, as well as the degree to which customer satisfaction and service quality predict customer purchase intention. The researcher used the *competing models strategy,* and proposed three possible relations among the variables. In model 1, satisfaction was proposed as an antecedent of service quality, and only service quality had a direct effect on purchase intention. In model 2, service quality was proposed as an antecedent of satisfaction, and only satisfaction had a direct effect on purchase intention. And in model 3, service quality and satisfaction were correlated, and both had a direct effect on purchase intention.

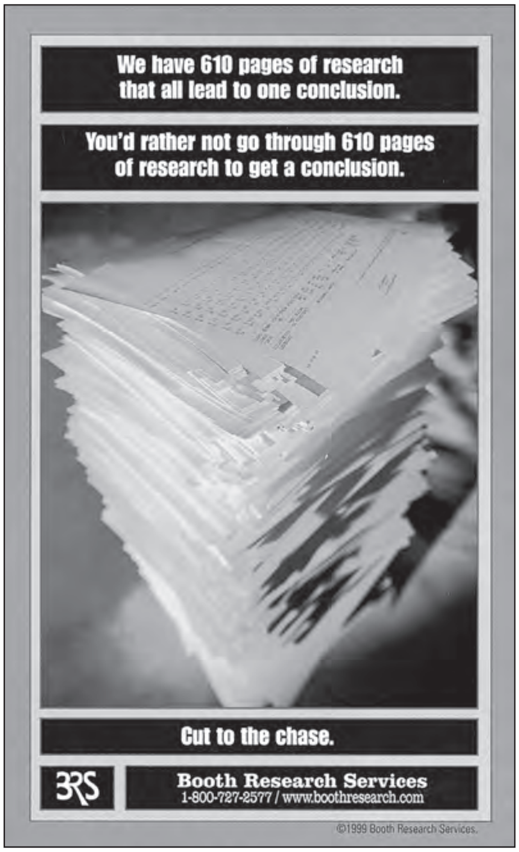
To collect the data, the researcher added three assumedly valid batteries of questions to the company's product and service warranty card. As soon as a large enough sample was obtained, the researcher specified the parameters of the proposed models and compared the implied structure with the covariance matrix of the data using maximum likelihood estimation as the iterative process.

The researcher finds that of the three proposed models, none of them have a satisfactory goodness of fit. However, of the three, model 2 seemed the most promising in that it yielded the lowest chi-square value and the highest value for the adjusted-goodness-of-fit index. After examining the second model's residual matrices and modification index, the researcher finds that the model could achieve a better fit if relation between service quality and purchase intention were not fixed. Accordingly, the researcher respecifies the model, freeing that parameter, and the implied matrix yields an acceptable goodness of fit. The implications of the results are that good service quality leads to customer satisfaction and that both variables have a direct effect on purchase intention (see Exhibit MV-10).

>Exhibit MV-10 Measurement Models Relative to the Full Structural Equation Model



The example in Exhibit MV-10 illustrates the three measurement models, one for each latent variable, relative to the full structural model. The three latent variables are satisfaction, service quality, and purchase intention, and each latent variable has three indicators. The direction of the single-pointed arrows from service quality and satisfaction to purchase intention denotes that purchase intention is a dependent variable in its relation to both service quality and satisfaction. However, although satisfaction is independent in its relation to purchase intention, it is dependent in its relation to service quality. The ability to model all three relations simultaneously is one of the foremost advantages of using SEM over other multivariate techniques.



As this Booth Research Services ad suggests, the researcher's role is to make sense of numerous data displays and thus assist the research sponsor in making an appropriate decision. Great data exploration and analysis will distill mountains of data printouts into insightful and supportable conclusions. **www.boothresearch.com**

Conjoint Analysis

The most common applications for conjoint analysis are market research and product development. Consumers buying a MindWriter computer, for example, may evaluate a set of attributes to choose the product that best meets their needs. They may consider brand, speed, price, educational value, games, or capacity for work-related tasks. The attributes and their features require that the buyer make trade-offs in the final decision making.

Method

Conjoint analysis typically uses input from nonmetric independent variables. Normally, we would use cross-classification tables to handle such data, but even multiway tables become quickly overwhelmed by the complexity. If there were three prices, three brands, three speeds, two levels of educational values, two categories for games, and two categories for work assistance, the model would have 216 decision levels ($3 \times 3 \times 3 \times 2 \times 2 \times 2$). A choice structure this size poses enormous difficulties for respondents and analysts. Conjoint analysis solves this problem with various optimal scaling approaches, often with loglinear models, to provide researchers with reliable answers that could not be obtained otherwise.

The objective of conjoint analysis is to secure **utility scores** (sometimes called *part-worths*) that represent the importance of each aspect of a product or service in the subjects' overall preference ratings. Utility scores are computed from the subjects' rankings or ratings of a set of cards. Each card in the deck describes one possible configuration of combined product attributes.

The first step in a conjoint study is to select the attributes most pertinent to the purchase decision. This may require an exploratory study such as a focus group, or it could be done by an expert with thorough market knowledge. The attributes selected are the independent variables, called *factors.* The possible values for an attribute are called *factor levels.* In the MindWriter example, the speed factor may have levels of 1.5 gigahertz and 3 gigahertz. Speed, like price, approaches linear measurement characteristics since consumers typically choose higher speeds and lower prices. Other factors like brand are measured as discrete variables.

After selecting the factors and their levels, a computer program determines the number of product descriptions necessary to estimate the utilities. SPSS procedures build a file structure for all possible combinations, generate the subset required for testing, produce the card descriptions, and analyze results. The command structure within these procedures provides for holdout sampling, simulations, and other requirements frequently used in commercial applications.[10]

Example

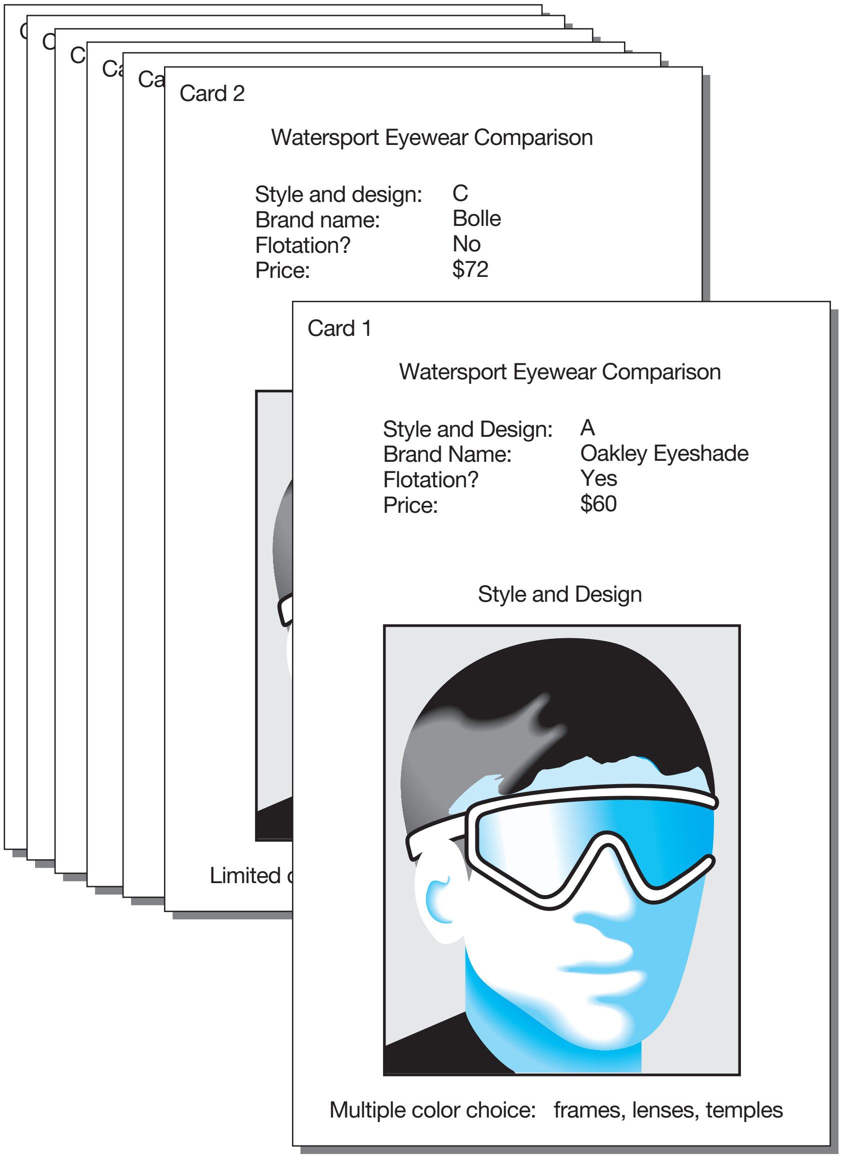
Watersports enthusiasts know the dangers of ultraviolet (UV) light. It fades paint and clothing; yellows surfboards, skis, and sailboards; and destroys sails. More important, UV damages the eye's retina and cornea. Americans spend more than $1.3 billion on 189 million pairs of sunglasses, most of which fail to provide adequate UV protection. Manufacturers of sunglasses for specialty markets have improved their products to such a degree that all of the companies in our example advertised 100 percent UV protection. Many other features influence trends in this market. For this example, we chose four factors from information contained in a review of sun protection products.[11]

Brand	Bolle	Hobbies	Oakley	Ski Optiks
Style*	A	A	A	A
	B	B	B	B
	C	C	C	C
Flotation	Yes	Yes	Yes	Yes
	No	No	No	No
Price	$100	$100	$100	$100
	$72	$72	$72	$72
	$60	$60	$60	$60
	$40	$40	$40	$40

*A = multiple color choices for frames, lenses, and temples.
B = multiple color choices for frames, lenses, and straps (no hard temples).
C = limited colors for frames, lenses, and temples.

This is a 4 (brand) × 3 (style) × 2 (flotation) × 4 (price) design, or a 96-option full-concept study. The algorithm selected 16 cards to estimate the utilities for the full concept. Combinations of interest that were not selected can be estimated later from the utilities. In addition, four holdout cards were

>Exhibit MV-11 Concept Cards for Conjoint Sunglasses Study



administered to subjects but evaluated separately. The cards shown in Exhibit MV-11 were administered to a small sample ($n = 10$). Subjects were asked to order their cards from most to least desirable. The data produced the results presented in Exhibits MV-12 and MV-13.

Exhibit MV-12 contains the results of the eighth participant's preferences. This individual was an avid windsurfer, and flotation was the most important attribute for her, followed by style and price and then brand. From her preferences, we can compute her maximum utility score:

$$\text{(Style B) } 3.46 + \text{(Oakley brand) } 1.31 + \text{(flotation) } 20.75 + \text{(price @ \$40) } 5.90 + \text{(constant) } - 8.21 = 23.21$$

If brand and price remain unchanged, a design that uses a hard temple with limited color choices (style C) and no flotation would produce a considerably lower total utility score for this respondent. For example:

$$\text{(Style C) } - 2.04 + \text{(Oakley brand) } 1.31 + \text{(no float) } 10.38 + \text{(price @ \$40) } 5.90 + \text{(constant) } - 8.21 = 7.34$$

>Exhibit MV-12 Conjoint Results for Participant 8, Sunglasses Study

Subject name: 8

Importance	Utility (s.e.)	Factor	Level *
23.86		STYLE	Style and design
	−1.4167(.3143)		A
	3.4583(.3685)		B
	−2.0417(.3685)		C
11.93		BRAND	Brand Name
	−1.4375(.4083)		Bolle
	.3125(.4083)		Hobbies
	1.3125(.4083)		Oakley
	−.1875(.4083)		Ski Optiks
45.01		FLOAT	Flotation?
	10.3750(.4715)		No
	20.7500(.9429)		Yes
	B = 10.3750(.4715)		
19.20		PRICE	Price *
	1.4750(.2108)		$100
	2.9500(.4217)		$72
	4.4250(.6325)		$60
	5.9000(.8434)		$40
	B = 1.4750(.2108)		
	−8.2083(.9163)	CONSTANT	

Pearson's r =	.994	Significance =	.0000
Pearson's r =	.990 for 4 holdouts	Significance =	.0051
Kendall's tau =	.967	Significance =	.0000
Kendall's tau =	1.000 for 4 holdouts	Significance =	.0208

*Subject reversed decision once.

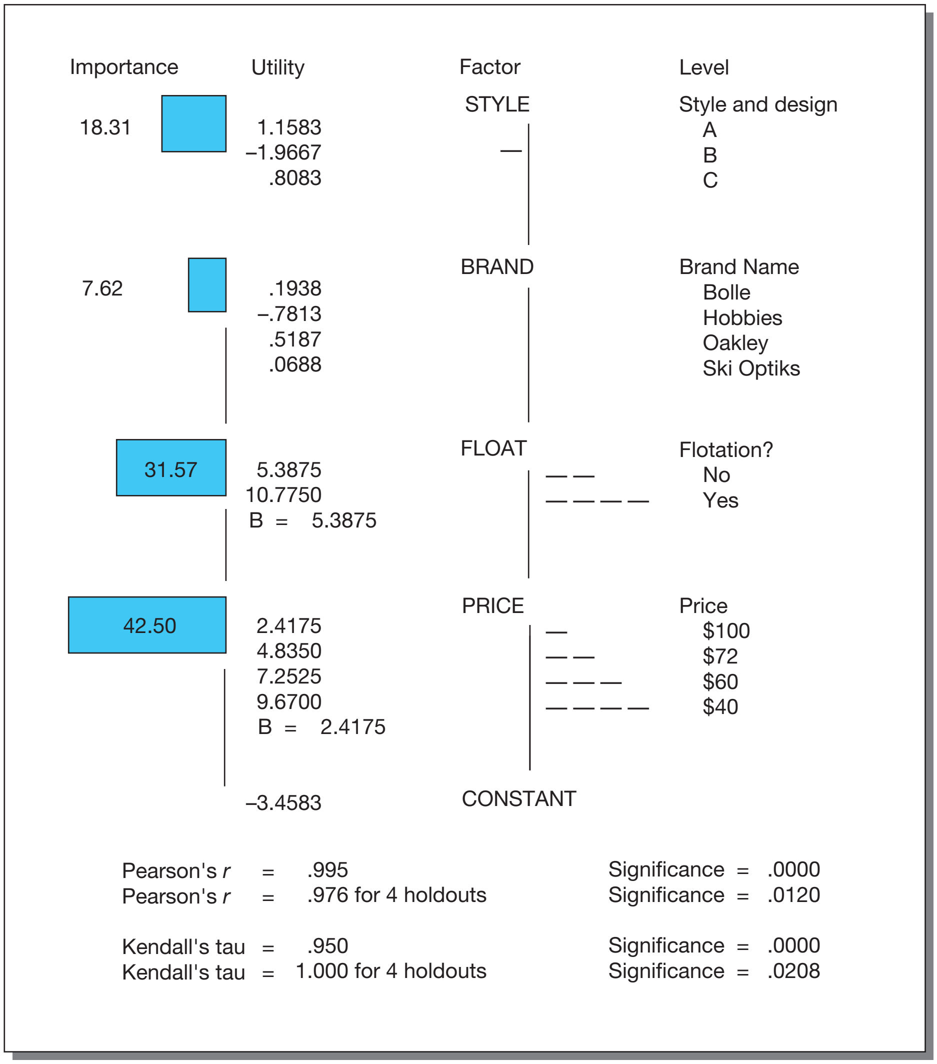
We could also calculate other combinations that would reveal the range of this individual's preferences. Our prediction that respondents would prefer less-expensive prices did not hold for the eighth respondent, as revealed by the asterisk next to the price factor in Exhibit MV-12. She reversed herself once on price to get flotation. Other subjects also reversed once on price to trade off for other factors.

The results for the sample are presented in Exhibit MV-13. In contrast to individuals, the sample placed price first in importance, followed by flotation, style, and brand. Group utilities may be calculated just as we did for the individual. At the bottom of the printout we find Pearson's r and Kendall's tau. Each was discussed in Chapter 18. In this application, they measure the relationship between observed and estimated preferences. Because holdout samples (in conjoint, regression, discriminant, and other methods) are not used to construct the estimating equation, the coefficients for the holdouts are often a more realistic index of the model's fit.

60 The percent of workers on four continents who trust their organization's senior leaders.

Conjoint analysis is an effective tool used by researchers to match preferences to known characteristics of market segments and design or target a product accordingly. See your student Online Learning Center for a MindWriter example of conjoint analysis using Simalto+Plus.

>Exhibit MV-13 Conjoint Results for Sunglasses Study Sample (n = 10)



> Interdependency Techniques

Factor Analysis

Unlike the predictor-criterion relationship present in dependence techniques, the variables used for interdependence techniques, like factor analysis, are not classified as dependent or independent. Because these variables are considered interrelated, they are analyzed simultaneously in order to identify an underlying structure. **Factor analysis** is a general term for several specific computational techniques used to examine patterns of relationships amongst select variables. The objective of these techniques is to create a more manageable number variables (data reduction) from a larger set of variables based on the nature and character of these relationships. For example, one may have data on 100 employees with scores on six attitude scale items.

Method

Factor analysis begins with the task of reducing the number of variables in order to simplify subsequent analyses. The data reduction process is based on the relationships or intercorrelations among the variables within the correlation matrix. Although this can be done in a number of ways, the most

>snapshot

The Mail as a "Super" E-Business

From 2000 to 2030, the United Nations estimates that the world's urban population will grow at 1.8 percent, resulting in an urban population that will double in 38 years. Urbanized populations increase demand for postal services and attract competition from nonpostal operators. With the world's postal system growing at a rate of 4 percent, according to the Universal Postal Union, hybrid mail will account for more than 6 percent or approximately 35 billion of the world's 550 billion pieces of physical mail by the mid-2000s.

SuperLetter.com is an e-business success story in the hybrid-mail sector. According to founder Christopher Schultheiss, "SuperLetter Hybrid Mail is secure, encrypted mail from point of origin to delivery to recipient, transported electronically to a print point closest to the addressee. Hybrid mail begins as a unique, serial-numbered encrypted data record and ends as a sealed letter delivered to the recipient. It is secure, tamper-proof and CBR-free from end to end." This is one of the reasons it has been successful with rapid mail delivery to the British Armed Forces and U.S. Marine Corps in Iraq, Afghanistan, and other parts of the world. As the world's first global "hybrid mail" network, it enables users to create letters, documents, and photos on their personal computers and send them like e-mail over the Internet to remote printers near the recipients, where they are printed, folded, enveloped, and delivered.

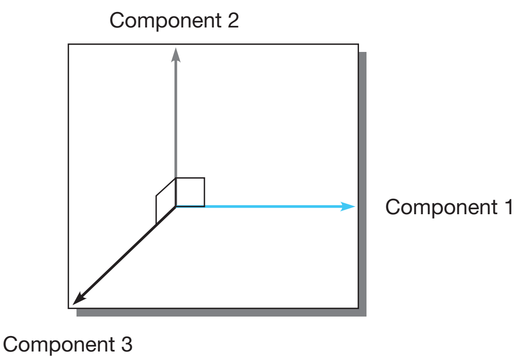
Using a variety of multivariate statistical techniques, SuperLetter identified target markets as government/military organizations, professional and financial service firms, not-for-profit organizations, and cruise ship and yacht mail. SuperLetter also draws business from the $100 billion worldwide international courier market, like FedEx, UPS, and DHL, now experiencing strong growth rates. By bridging the gap between conventional door-to-door postal services, which take from 5 to10 days for overseas delivery, and private express/courier services, which may take from two to three days, SuperLetter's basic international service delivers a letter from desk to door in two to three days for about one-tenth of private courier costs and under one-half of the courier costs for same-day service.

www.superletter.com

frequently used approach is **principal components analysis.** This method transforms a set of variables into either a smaller number of variables that represent those in the original set or a completely new set of composite variables or principal components that are not correlated with each other. These linear combinations of variables, called **factors,** account for the variance in the data as a whole. The first principal component, or first factor, is comprised of the best linear function of the original variables as to maximize the amount of the total variance that can be explained. The second principal component is defined as the best linear combination of variables for explaining the variance *not* accounted for by the first factor. In turn, there may be a third, fourth, and *k*th component, each being the best linear combination of variables not accounted for by the previous factors.

The process continues until all the variance is accounted for, but as a practical matter, it is usually stopped after a small number of factors have been extracted. It is important to note that principal components, or factors, will always be produced in a factor analysis. However, the quality and usefulness of the derived factors are dependent upon the types, number, and conceptual basis of variables selected for inclusion during the initial research design. The output of a principal components analysis might look like the hypothetical data shown in Exhibit MV-14.

>Exhibit MV-14 Principal Components Analysis from a Three-Variable Data Set



Extracted Components	% of Variance Accounted For	Cumulative Variance
Component no. 1	63%	63%
Component no. 2	29%	92%
Component no. 3	8%	100%

>Exhibit MV-15 Factor Matrices

	A Unrotated Factors			B Rotated Factors	
Variable	I	II	*h*2	I	II
A	0.70	−.40	0.65	0.79	0.15
B	0.60	−.50	0.61	0.75	0.03
C	0.60	−.35	0.48	0.68	0.10
D	0.50	0.50	0.50	0.06	0.70
E	0.60	0.50	0.61	0.13	0.77
F	0.60	0.60	0.72	0.07	0.85
Eigenvalue	2.18	1.39			
Percent of variance	36.3	23.2			
Cumulative percent	36.3	59.5			

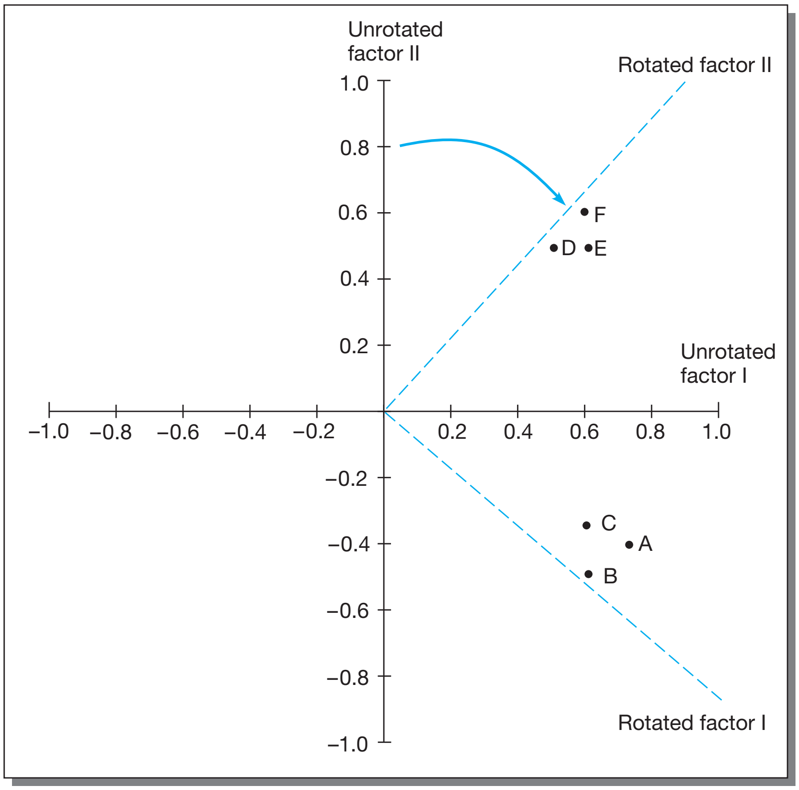
Numerical results from a factor study are shown in Exhibit MV-15. The values in this table are correlation coefficients between the factor and the variables (.70 is the *r* between variable A and factor I). These correlation coefficients are called **loadings.** The higher the loading, the greater the contribution of the variable in defining the particular factor. Two other elements in Exhibit MV-15 need explanation. **Eigenvalues** are the sum of the variances of the factor values (for factor I the eigenvalue is $.70^2 + .60^2 + .50^2 + .60^2 + .60^2$). When divided by the number of variables, an eigenvalue yields an estimate of the amount of total variance explained by the factor. For example, factor I accounts for 36 percent of the total variance. If a factor has a low eigenvalue, then it adds little to the explanation of variances in the variables and may be disregarded. The column headed "h^2" gives the **communalities,** or estimates of the variance in each variable that is explained by the two factors. With variable A, for example, the communality is $.70^2 + (-.40)^2 = .65$, indicating that 65 percent of the variance in variable A is statistically explained in terms of factors I and II.

As displayed in column A, the unrotated factor loadings are not informative because there are no definitive "drivers" for either of the two factors within the current structure. What one would like to find is some pattern in which factor I would be heavily loaded (have a high *r*) on some variables and factor II on others. Such a condition would suggest rather "pure" constructs underlying each factor. You attempt to secure this less ambiguous condition between factors and variables by **rotation.** This procedure allows choices between orthogonal and oblique methods. (When the factors are intentionally rotated to result in no correlation between the factors in the final solution, this procedure is called *orthogonal*; when the factors are not manipulated to be zero correlation but may reveal the degree of correlation that exists naturally, it is called *oblique.*) We illustrate an orthogonal solution here.

To understand the rotation concept, consider that you are dealing only with simple two-dimensional rather than multidimensional space. The variables in Exhibit MV-15 can be plotted in two-dimensional space as shown in Exhibit MV-16. Two axes divide this space, and the points are positioned relative to these axes. The location of these axes is arbitrary, and they represent only one of an infinite number of reference frames that could be used to reproduce the matrix. As long as you do not change the intersection points and keep the axes at right angles, when an orthogonal method is used, you can rotate the axes to find a better solution or position for the reference axes. "Better" in this case means a matrix that makes the factors as pure as possible (each variable loads onto as few factors as possible to simplify the structure). After performing the rotation as shown in Exhibit MV-16, the factor solution is improved substantially. Returning to Exhibit MV-15, column B, Rotated Factors, suggests that the measurements from six scales may be summarized by two underlying factors. Variables A, B, and C load primarily on factor I, whereas variables D, E, and F load heavily on factor II.

In this hypothetical example, the rotation resulted in the variables loading unambiguously onto only one of the two derived factors. In practice, however, the results displayed in the factor matrix are not always as clear, prompting the researcher to consider changes to the factor model such as using another rotation method, selecting a solution with either a greater or lesser number of factors, or deleting

>Exhibit MV-16 Orthogonal Factor Rotations



variables that lack significant loadings on any of the factors. Once this process is complete, the ultimate interpretation of factor loadings is largely subjective. There is no way to calculate the meanings of factors; they are what one sees in them, taking into consideration, as noted previously, the nature and conceptual basis of the variables included in the analysis. For this reason, factor analysis is largely used for exploration, allowing one to detect patterns in latent variables, discover new concepts, and reduce data.

Example

Student grades make an interesting example to illustrate the use of factor analysis. The chairperson of Metro U's MBA program has been reviewing grades for the first-year students and is struck by the patterns in the data. His hunch is that distinct types of people are involved in the study of business, and he decides to gather evidence for this idea.

Suppose the chairperson selects a sample of 21 grade reports from 10 of the program's courses for first-year students whose GPAs fall within the middle range. His analysis involves the following three steps:

1. Calculate a correlation matrix between the grades for all pairs of the 10 courses for which data exist.
2. Factor-analyze the matrix by the principal components method.
3. Select a rotation procedure to clarify the factors and aid in interpretation.

Exhibit MV-17 shows a portion of the correlation matrix. These data represent correlation coefficients between the 10 courses. For example, grades secured in V1 (Financial Accounting) correlated rather well (0.56) with grades received in course V2 (Managerial Accounting). The next best correlation with V1 grades is an inverse correlation (−.44) with grades in V7 (Production).

After the calculation of correlation matrix, the extraction of components is shown in Exhibit MV-18. Although the program will produce a table with as many as 10 factors, you choose, in this case, to stop

>Exhibit MV-17 Correlation Coefficients, Metro U MBA Study

Variable	Course	V1	V2	V3	V10
V1	Financial Accounting	1.00	0.56	0.17	−.01
V2	Managerial Accounting	0.56	1.00	−.22	0.06
V3	Finance	0.17	−.22	1.00	0.42
V4	Marketing	−.14	0.05	−.48	−.10
V5	Human Behavior	−.19	−.26	−.05	−.23
V6	Organization Design	−.21	−.00	−.56	−.05
V7	Production	−.44	−.11	−.04	−.08
V8	Probability	0.30	0.06	0.07	−.10
V9	Statistical Inference	−.05	0.06	−.32	0.06
V10	Quantitative Analysis	−.01	0.06	0.42	1.00

>Exhibit MV-18 Factor Matrix Using Principal Factor with Iterations, Metro U MBA Study

Variable	Course	Factor 1	Factor 2	Factor 3	Communality
V1	Financial Accounting	0.41	0.71	0.23	0.73
V2	Managerial Accounting	0.01	0.53	−.16	0.31
V3	Finance	0.89	−.17	0.37	0.95
V4	Marketing	−.60	0.21	0.30	0.49
V5	Human Behavior	0.02	−.24	−.22	0.11
V6	Organization Design	−.43	−.09	−.36	0.32
V7	Production	−.11	−.58	−.03	0.35
V8	Probability	0.25	0.25	−.31	0.22
V9	Statistical Inference	−.43	0.43	0.50	0.62
V10	Quantitative Analysis	0.25	0.04	0.35	0.19
Eigenvalue		1.83	1.52	0.95	
Percent of variance		18.30	15.20	9.50	
Cumulative percent		18.30	33.50	43.00	

the process after three factors have been extracted. Several features in this table are worth noting. Recall that the communalities indicate the amount of variance in each variable that is being "explained" by the factors. Thus, these three factors account for about 73 percent of the variance in grades in the financial accounting course. It should be apparent from these communality figures that some of the courses are not explained well by the factors selected.

The eigenvalue row in Exhibit MV-18 is a measure of the explanatory power of each factor. For example, the eigenvalue for factor 1 is 1.83 and is computed as follows:

$$1.83 = (.41)^2 + (.01)^2 + \cdots + (.25)^2$$

The percent of variance accounted for by each factor in Exhibit MV-18 is computed by dividing eigenvalues by the number of variables. When this is done, one sees that the three factors account for about 43 percent of the total variance in course grades.

>Exhibit MV-19 Varimax Rotated Factor Matrix, Metro U MBA Study

Variable	Course	Factor 1	Factor 2	Factor 3
V1	Financial Accounting	0.84	0.16	−.06
V2	Managerial Accounting	0.53	−.10	0.14
V3	Finance	−.01	0.90	−.37
V4	Marketing	−.11	−.24	0.65
V5	Human Behavior	−.13	−.14	−.27
V6	Organization Design	−.08	−.56	−.02
V7	Production	−.54	−.11	−.22
V8	Probability	0.41	−.02	−.24
V9	Statistical Inference	0.07	0.02	0.79
V10	Quantitative Analysis	−.02	0.42	0.09

In an effort to further clarify the factors, a varimax (orthogonal) rotation is used to secure the matrix shown in Exhibit MV-19. The largest factor loadings for the three factors are as follows:

Factor 1		Factor 2		Factor 3	
Financial Accounting	0.84	Finance	0.90	Marketing	0.65
Managerial Accounting	0.53	Organization Design	−.56	Statistical Inference	0.79
Production	−.54				

Interpretation

The varimax rotation appears to clarify the relationship among course grades, but as pointed out earlier, the interpretation of the results is largely subjective. We might interpret the above results as showing three kinds of students, classified as the accounting, finance, and marketing types.

However, a number of problems may affect the interpretation of these results. Among the major concerns are:

1. The sample is small and any attempt at replication, such as using a split sample, might produce a different pattern of factor loadings. Factor stability is affected by both the sample size and number of cases per variable.
2. Altering your decision to extract a different number of factors (rather than three) can result in variables loading to different degrees or onto different factors.
3. Even if the findings are replicated, the variance in the data may be due to the varying influence of professors or the way they teach the courses, rather than to the subject content or types of students.
4. The labels or classifications may not truly reflect the latent construct that underlies any factors we extract.

Factor analysis is a powerful technique that can be used to identify underlying dimensions for a set of variables. However, as this simple example illustrates, issues regarding research design, derived factors, and interpretation must be addressed with great care.

Cluster Analysis

Unlike techniques for analyzing the relationships between variables, **cluster analysis** is a set of interdependence techniques for grouping similar objects or people. Originally developed as a classification device for taxonomy, its use has spread because of classification work in medicine, biology, and other sciences. Its visibility in those fields and the availability of high-speed computers to carry out the extensive calculations have sped its adoption in business. Understanding one's market very often involves segmenting customers into homogeneous groups that have common buying characteristics or behave in similar ways. Such segments frequently share similar psychological, demographic, lifestyle, age, financial, or other characteristics.

Cluster analysis offers a means for segmentation research and other business problems, such as understanding buyer behaviors, where the goal is to identify similar groups. It shares some similarities with factor analysis, especially when factor analysis is applied to people (*Q*-analysis) instead of to variables. However, cluster analysis treats correlations as similarity (distance) measures whereas *Q*-analysis takes into account control variables (as in a general linear model). It differs from discriminant analysis in that discriminant analysis begins with a well-defined group composed of two or more distinct sets of characteristics in search of a set of variables to separate them. In contrast, cluster analysis starts with an undifferentiated group of people, events, or objects and attempts to reorganize them into homogeneous subgroups.

Method

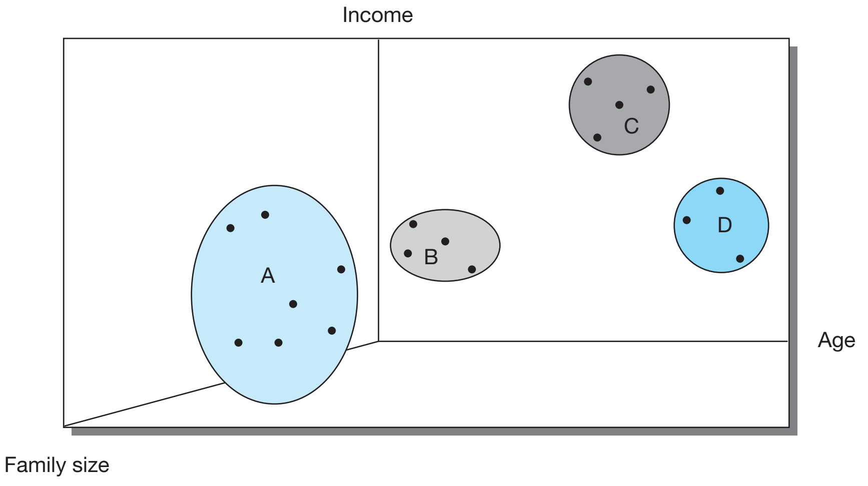
Five steps are basic to the application of most cluster studies:

1. Selection of the sample to be clustered (e.g., buyers, medical patients, inventory, products, employees).
2. Definition of the variables on which to measure the objects, events, or people (e.g., market segment characteristics, product competition definitions, financial status, political affiliation, symptom classes, productivity attributes).
3. Computation of similarities among the entities through correlation, Euclidean distances, and other techniques.
4. Selection of mutually exclusive clusters (maximization of within-cluster similarity and between-cluster differences) or hierarchically arranged clusters.
5. Cluster comparison and validation.

Different clustering methods can and do produce different solutions. It is important to have enough information about the data, as well as the various clustering algorithms and stopping rules, to know when the derived groups are real and not merely imposed on the data by the method. In addition, issues such as variable scaling and intercorrelations can influence the final cluster solution.

The example in Exhibit MV-20 shows a cluster analysis of individuals based on three dimensions: age, income, and family size. Cluster analysis could be used to segment the car-buying population into distinct markets. For example, cluster A might be targeted as potential minivan or sport-utility vehicle

>Exhibit MV-20 Cluster Analysis on Three Dimensions



buyers. The market segment represented by cluster B might be a sports and performance car segment. Clusters C and D could both be targeted as buyers of sedans, but the C cluster might be the luxury buyer. This form of clustering or a hierarchical arrangement of the clusters may be used to plan marketing campaigns and develop strategies.

Example

The entertainment industry is a complex business. A huge number of films are released each year internationally with some notable financial surprises. Paris offers one of the world's best selections of films and sources of critical review for predicting an international audience's acceptance. Residents of New York and Los Angeles are often surprised to discover their cities are eclipsed by Paris's average of 300 films per week shown in over 100 locations.

We selected ratings from 12 cinema reviewers using sources ranging from *Le Monde* to international publications sold in Paris. The reviews reputedly influence box-office receipts, and the entertainment business takes them seriously.

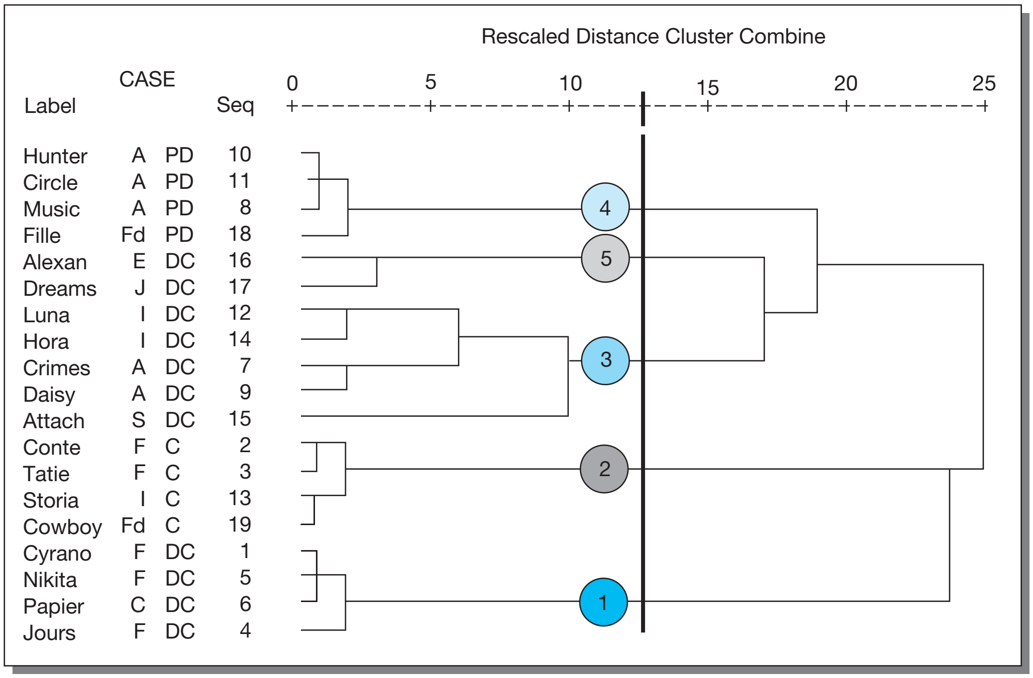
The object of this cluster example was to classify 19 films into homogeneous subgroups. The production companies were American, Canadian, French, Italian, Spanish, Finnish, Egyptian, and Japanese. Three genres of film were represented: comedy, dramatic comedy, and psychological drama. Exhibit MV-21 shows the data by film name, country of origin, and genre. The table also lists the clusters for each film using the **average linkage method.** This approach considers distances between all possible pairs rather than just the nearest or farthest neighbor.

The sequential development of the clusters and their relative distances are displayed in a diagram called a *dendogram.* Exhibit MV-22 shows that the clustering agglomerative procedure begins with 19 films and continues until all the films are again an undifferentiated group. The solid vertical line shows the point

>Exhibit MV-21 Film, Country, Genre, and Cluster Membership

				Number of Clusters			
Film	**Country**	**Genre**	**Case**	**5**	***4***	***3***	***2***
Cyrano de Bergerac	France	DramaCom	1	1	1	1	1
Il y a des Jours	France	DramaCom	4	1	1	1	1
Nikita	France	DramaCom	5	1	1	1	1
Les Noces de Papier	Canada	DramaCom	6	1	1	1	1
Leningrad Cowboys . . .	Finland	Comedy	19	2	2	2	2
Storia de Ragazzi . . .	Italy	Comedy	13	2	2	2	2
Conte de Printemps	France	Comedy	2	2	2	2	2
Tatie Danielle	France	Comedy	3	2	2	2	2
Crimes and Misdem . . .	USA	DramaCom	7	3	3	3	2
Driving Miss Daisy	USA	DramaCom	9	3	3	3	2
La Voce della Luna	Italy	DramaCom	12	3	3	3	2
Che Hora E	Italy	DramaCom	14	3	3	3	2
Attache-Moi	Spain	DramaCom	15	3	3	3	2
White Hunter Black . . .	USA	PsyDrama	10	4	4	3	2
Music Box	USA	PsyDrama	8	4	4	3	2
Dead Poets Society	USA	PsyDrama	11	4	4	3	2
La Fille aux All . . .	Finland	PsyDrama	18	4	4	3	2
Alexandrie, Encore . . .	Egypt	DramaCom	16	5	3	3	2
Dreams	Japan	DramaCom	17	5	3	3	2

>Exhibit MV-22 Dendogram of Film Study Using Average Linkage Method



at which the clustering solution best represents the data. This determination was guided by coefficients provided by the SPSS program for each stage of the procedure. Five clusters explain this data set.

The first cluster shown in Exhibit MV-22 has three French-language films and one Canadian film, all of which are dramatic comedies. Cluster 2 consists of comedy films. Two French and two other European films joined at the first stage, and then these two groups came together at the second stage. Cluster 3, composed of dramatic comedies, is otherwise diverse. It is made up of two American films with two Italian films adding to the group at the fourth stage. Late in the clustering process, cluster 3 is completed when a Spanish film is appended. In cluster 4, we find three American psychological dramas combined with a Finnish film at the second stage. In cluster 5, two very different dramatic comedies are joined in the third stage.

Cluster analysis classified these productions based on reviewers' ratings. The similarities and distances are influenced by film genre and culture (as defined by the translated language). Comparison of the derived clusters reveals their particular characteristics, allowing the researcher to assign a descriptive label to each cluster. More importantly however, this examination assists in determining how the final cluster solution compares with those suggested by either prior research or by practical experience.

Because the optimal number of clusters formed is often at the discretion of the researcher, assessing the validity of the cluster solution is important to ensure that the process and methods are not inadvertently misapplied. There are several approaches that assist in validating the clusters, including utilizing a separate sample, splitting the existing sample into two groups, or selecting variables not included in the original analysis but known from prior research to vary across the clusters. For each of these methods, the results of cluster analysis are compared with the original findings.

Multidimensional Scaling

Multidimensional scaling (MDS) creates a special description of a respondent's perception about a product, service, or other object of interest on a *perceptual map.* This often helps the researcher to understand difficult-to-measure constructs such as product quality or desirability. In contrast to variables that can be measured directly, many constructs are perceived and cognitively mapped in different ways by individuals. With MDS, items that are perceived to be similar will fall close together on the perceptual map, and items that are perceived to be dissimilar will be farther apart.

Method

We may think of three types of attribute space, each representing a multidimensional map. First, there is *objective space,* in which an object can be positioned in terms of its measurable attributes: its flavor, weight, and nutritional value. Second, there is *subjective space,* where perceptions of the object's flavor, weight, and nutritional value may be positioned. Objective and subjective attribute assessments may coincide, but often they do not. A comparison of the two allows us to judge how accurately an object is being perceived. Individuals may hold different perceptions of an object simultaneously, and these may be averaged to present a summary measure of perceptions. In addition, a person's perceptions may vary over time and in different circumstances; such measurements are valuable to gauge the impact of various perception-affecting actions, such as advertising programs.

With a third map we can describe respondents' preferences using the object's attributes. This represents their ideal; all objects close to this ideal point are interpreted as preferred by respondents to those that are more distant. Ideal points from many people can be positioned in this preference space to reveal the pattern and size of preference clusters. These can be compared to the subjective space to assess how well the preferences correspond to perception clusters. In this way, cluster analysis and MDS can be combined to map market segments and then examine products designed for those segments.

Example

We illustrate multidimensional scaling with a study of 16 restaurants in a resort area.[12] The restaurants chosen represent medium-price family restaurants to high-price gourmet restaurants. We created a metric algorithm measuring the similarities among the 16 restaurants by asking patrons questions on a 5-point metric scale about different dimensions of service quality and price. The matrix of similarities is shown in Exhibit MV-23. Higher numbers reflect the items that are more dissimilar.

We might also ask participants to judge the similarities between all possible pairs of restaurants; then we produce a matrix of similarities using (nonmetric) ordinal data. The matrix would contain ranks with 1 representing the most similar pair and *n* indicating the most dissimilar pair.

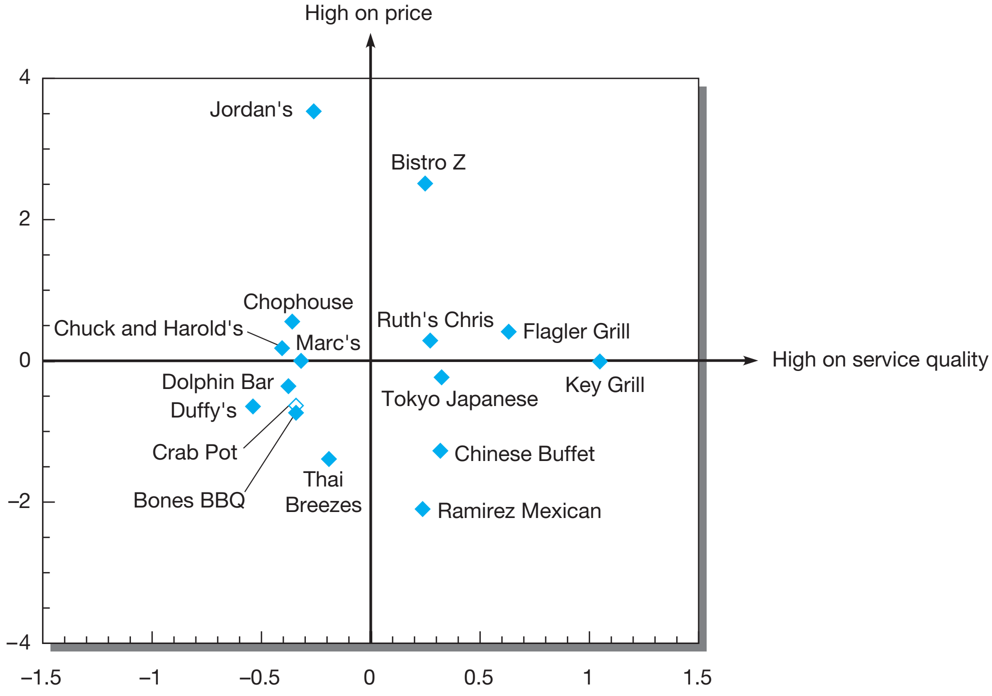
A computer program is used to analyze the data matrix and generate a perceptual map.[13] The objective is to find a multidimensional spatial pattern that best reproduces the original order of the data. For example, the most similar pair (restaurants 3, 6) must be located in this multidimensional space closer together than any other pair. The least similar pair (restaurants 14, 15) must be the farthest apart. The computer program presents these relationships as a geometric configuration so that all distances between pairs of points closely correspond to the original matrix.

Determining how many dimensions to use is complex. The more dimensions of space we use, the more likely the results will closely match the input data. Any set of *n* points can be satisfied by a

>Exhibit MV-23 Similarities Matrix of 16 Restaurants

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0															
2	3.9	0														
3	4.7	6.7	0													
4	4.4	2.8	4.7	0												
5	14.0	12.4	18.5	15.2	0											
6	4.9	6.9	0.2	4.9	18.7	0										
7	0.8	3.7	4.1	3.7	14.5	4.3	0									
8	6.0	2.1	8.5	4.0	11.8	8.7	5.8	0								
9	4.3	6.9	1.1	5.3	18.3	1.2	3.8	8.9	0							
10	8.2	4.9	8.5	4.1	15.3	8.6	7.6	3.9	9.3	0						
11	8.6	8.7	4.7	5.9	21.1	4.5	7.8	9.7	5.7	7.7	0					
12	2.2	3.7	6.9	5.5	11.8	7.1	2.8	5.5	6.5	8.5	10.5	0				
13	8.4	9.8	3.7	7.2	22.0	3.5	7.8	11.2	4.5	10.0	2.9	10.6	0			
14	12.8	13.4	8.2	10.6	25.8	8.1	12.1	14.4	9.1	12.0	4.7	14.9	4.6	0		
15	19.1	18.2	23.8	21.0	6.2	24.0	19.7	17.8	23.4	21.5	26.9	16.9	27.4	31.5	0	
16	2.6	5.2	2.1	4.0	16.5	2.3	2.0	7.2	1.9	8.0	6.3	4.8	5.8	10.3	21.7	0

>Exhibit MV-24 Positioning of Selected Restaurants



configuration of $n - 1$ dimensions. Our aim, however, is to secure a structure that provides a good fit for the data and has the fewest dimensions. MDS is best understood using two or at most three dimensions.

Most software programs include the calculation of a **stress index** (S-stress or Kruskal's stress) that ranges from the worst fit (1) to the perfect fit (0). This study, for example, had a stress of .001. Another index, R^2, is interpreted as the proportion of variance of transformed data accounted for by distances in the model. A result close to 1.0 is desirable.

In the restaurants example, we conclude that two dimensions represent an acceptable geometric configuration, as shown in Exhibit MV-24. The distance between Crab Pot and Bones BBQ (3, 6) is the shortest, while that between Ramirez Mexican and Jordan's (14, 15) is the longest. As with factor analysis, there is no statistical solution to the definition of the dimensions represented by the X and Y axes. The labeling is judgmental and depends on the insight of the researcher, analysis of information collected from respondents, or another basis. Respondents sometimes are asked to state the criteria they used for judging the similarities, or they are asked to judge a specific set of criteria.

Consistent with raw data, Jordan's and Bistro Z have high price but service quality close to the sample mean. In contrast, Flagler and Key Grills generated a price close to the sample's average while providing higher service quality. We could hypothesize that the latter two restaurants may be run more efficiently—are smaller and less complex—but that would need to be confirmed with another study. The clustering of companies in attribute space shows that they are perceived to be similar along the dimensions measured.

MDS is most often used to assess perceived similarities and differences among objects. Using MDS allows the researcher to understand constructs that are not directly measurable. The process provides a spatial map that shows similarities in terms of relative distances. It is best understood when limited to two or three dimensions that can be graphically displayed.

Exhibit MV-25 concludes our coverage of multivariate analysis with sample articles representing each technique described.

>Exhibit MV-25 Research Articles that Illustrate Each Multivariate Technique

Multivariate Method	Journal Article References
Multiple regression	Baker, T., Cronin, J. & Hopkins, C. (2009). The impact of involvement on key service relationships. *Journal of Services Marketing, 23*(2), 115–124. Lambrecht, K., Kaefer, F. & Ramenofsky, S. (2009). Sportscape factors influencing spectator attendance and satisfaction at a professional golf association tournament. *Sport Marketing Quarterly, 18*(3), 165–172.
Discriminant analysis	Li, P. (2010). Effects of individual differences on choice strategy in goal-directed online shopping. *The Journal of American Academy of Business, 15*(2), 186–192. Kuruvilla, S., Joshi, N. & Shah, N. (2009). Do men and women really shop differently? An exploration of gender differences in mall shopping in India. *International Journal of Consumer Studies, 33*(6), 715–723.
MANOVA	Meyer-Waarden, L. & Benavent, C. (2009). Grocery retail loyalty program effects: Self-selection or purchase behavior change? *Journal of the Academy of Marketing Sciences, 37*(3), 345–358. Smith, S. & Costello, C. (2009). Culinary tourism: Satisfaction with a culinary event utilizing importance-performance grid analysis. *Journal of Vacation Marketing, 15*(2), 99–110.
SEM	Kilic, C. & Dursun, T. (2010). The effect of organizational culture on customer orientation. *Journal of American Academy of Business, 15*(2), 1–7. Botonaki, A., Natos, D. & Konstadinos, M. (2009). Exploring convenience food consumption through a structural equation model. *Journal of Food Products Marketing, 15*(1), 64–79.
Conjoint analysis	Rokka, J. & Uusitalo, L. (2009). Preference for green packaging in consumer product choices—Do consumers care? *International Journal of Consumer Studies, 32*(5), 516–525. Han, H. (2010). The investigation of country-of-origin effect-using Taiwanese consumers' perceptions of luxury handbags as example. *Journal of American Academy of Business, 15*(2), 66–72.
Factor analysis	Krystallis, A. & Chryssochoidis, G. (2009). Does the country of origin (COO) of food products influence consumer evaluations? An empirical examination of ham and cheese. *Journal of Food Products Marketing, 15*(3), 283–303. Lu, C. & Yang, C. (2010). Logistics service capabilities and firm performance of international distribution center operators. *The Service Industries Journal, 30*(2), 281–298.
Cluster analysis	Thomas, H. & Li, X. (2009). Mapping globally branded business schools: A strategic positioning analysis. *Management Decision, 47*(9), 1420–1440. Boo, S. & Jones, D. (2009). Using a validation process to develop market segmentation based on travel motivation for major metropolitan areas. *Journal of Travel & Tourism Marketing, 26*(1), 60–79.
Multidimensional scaling (MDS)	Bao, J. & Sweeney, J. (2009). Comparing factor analytical and circumplex models of brand personality in brand positioning. *Psychology & Marketing, 26*(10), 927–949. Jin, Y. & Kelsay, C. (2008). Typology and dimensionality of litigation public relations strategies: The Hewlett-Packard board pretexting scandal case. *Public Relations Review, 34*(1), 66–69.

>summary

1 Multivariate techniques are classified into two categories: dependency and interdependency. When a problem reveals the presence of criterion and predictor variables, we have an assumption of dependence. If the variables are interrelated without designating some as dependent and others independent, then interdependence of the variables is assumed. The choice of techniques is guided by the number of dependent and independent variables involved and whether they are measured on metric or nonmetric scales.

2 Multiple regression is an extension of bivariate linear regression. When a researcher is interested in explaining or predicting a metric dependent variable from a set of metric independent variables (although dummy variables may also be used), multiple regression is often selected. Regression results provide information on the statistical significance of the independent variables, the strength of association between one or more of the predictors and the criterion, and a predictive equation for future use.

3 Discriminant analysis is used to classify people or objects into groups based on several predictor variables. The groups are defined by a categorical variable with two or more values, whereas the predictors are metric. The

effectiveness of the discriminant equation is based not only on its statistical significance but also on its success in correctly classifying cases to groups.

4 Multivariate analysis of variance, or MANOVA, is one of the more adaptive techniques for multivariate data. MANOVA assesses the relationship between two or more metric dependent variables and classificatory variables or factors. MANOVA is most commonly used to test differences among samples of people or objects. In contrast to ANOVA, MANOVA handles multiple dependent variables, thereby simultaneously testing all the variables and their interrelationships.

5 Researchers have relied increasingly on structural equation modeling (SEM) to test hypotheses about the dimensionality of, and relationships among, latent and observed variables. Researchers refer to structural equation models as LISREL (linear structural relations) models. The major advantages of SEM are (1) that multiple and interrelated dependence relationships can be estimated simultaneously and (2) that it can represent unobserved concepts, or latent variables, in these relationships and account for measurement error in the estimation process. Researchers using SEM must follow five basic steps: (1) model specification, (2) estimation, (3) evaluation of fit, (4) respecification of the model, and (5) interpretation and communication.

6 Conjoint analysis is a technique that typically handles nonmetric independent variables. Conjoint analysis allows the researcher to determine the importance of product or service attributes and the levels or features that are most desirable. Respondents provide preference data by ranking or rating cards that describe products. These data become utility weights of product characteristics by means of optimal scaling and loglinear algorithms.

7 Principal components analysis extracts uncorrelated factors that account for the largest portion of variance from an initial set of variables. Factor analysis also attempts to reduce the number of variables and discover the underlying constructs that explain the variance. A correlation matrix is used to derive a factor matrix from which the best linear combination of variables may be extracted. In many applications, the factor matrix will be rotated to simplify the factor structure.

8 Unlike techniques for analyzing the relationships between variables, cluster analysis is a set of techniques for grouping similar objects or people. The cluster procedure starts with an undifferentiated group of people, events, or objects and attempts to reorganize them into homogeneous subgroups using a set of variables as the basis for their similarity.

9 Multidimensional scaling (MDS) is often used in conjunction with cluster analysis or conjoint analysis. It allows a respondent's perception about a product, service, or other object of attitude to be described in a spatial manner. MDS helps the business researcher to understand difficult-to-measure constructs such as product quality or desirability, which are perceived and cognitively mapped in different ways by different individuals. Items judged to be similar will fall close together in multidimensional space and are revealed numerically and geometrically by spatial maps.

>**key**terms

>discussionquestions

Terms in Review

1 Distinguish among multidimensional scaling, cluster analysis, and factor analysis.

2 Describe the differences between dependency techniques and interdependency techniques. When would you choose a dependency technique?

Making Research Decisions

3 How could discriminant analysis be used to provide insight into MANOVA results where the MANOVA has one independent variable (a factor with two levels)?

4 Describe how you would create a conjoint analysis study of off-road vehicles. Restrict your brands to three, and suggest possible factors and levels. The full-concept description should not exceed 256 decision options.

5 What type of multivariate method do you recommend in each of the following cases and why?

   a You want to develop an estimating equation that will be used to predict which applicants will come to your university as students.

   b You would like to predict family income using such variables as education and stage in family life cycle.

   c You wish to estimate standard labor costs for manufacturing a new dress design.

   d You have been studying a group of successful salespeople. You have given them a number of psychological tests. You want to bring meaning out of these test results.

6 Sales of a product are influenced by the salesperson's level of education and gender, as well as consumer income, ethnicity, and wealth.

   a Formulate this statement as a multiple regression model (form only, without parameter estimation).

   b Specify dummy variables.

   c If the effects of consumer income and wealth are not additive alone, and an interaction is expected, specify a new variable to test for the interaction.

7 What multivariate technique would you use to analyze each of the following problems? Explain your choice.

   a Employee job satisfaction (high, normal, low) and employee success (0–2 promotions, 3–5 promotions, 5+ promotions) are to be studied in three different departments of a company.

   b Consumers making a brand choice decision among three brands of coffee are influenced by their own income levels and the extent of advertising of the brands.

   c Consumer choice of color in fabrics is largely dependent on ethnicity, income levels, and the temperature of the geographic area. There is detailed areawide demographic data on income levels, ethnicity, and population, as well as the weather bureau's historical data on temperature. How would you identify geographic areas for selling dark-colored fabric? You have sample data for 200 randomly selected consumers: their fabric color choice, income, ethnicity, and the average temperature of the area where they live.

From Concept to Practice

8 An analyst sought to predict the annual sales for a home-furnishing manufacturer using the following predictor variables:

X_1 = marriages during the year

X_2 = housing starts during the year

X_3 = annual disposable personal income

X_4 = time trend (first year = 1, second year = 2, and so forth)

Using data for 24 years, the analyst calculated the following estimating equation:

$$Y = 49.85 - .068X_1 + .036X_2 + 1.22X_3 - 19.54X_4$$

The analyst also calculated an $R^2 = .92$ and a standard error of estimate of 11.9. Interpret the above equation and statistics.

9 You are working with a consulting group that has a new project for the Palm Grove School System. The school system of this large county has individuals with purchasing, service, and maintenance responsibilities. They were asked to evaluate the vendor/distribution channels of products that the county purchases. The evaluations were on a 10-point metric scale for the following variables:

Delivery speed—amount of time for delivery once the order has been confirmed.

Price level—level of price charged by the product suppliers.

Price flexibility—perceived willingness to negotiate on price.

Manufacturer's image—manufacturer or supplier's image.

Overall service—level of service necessary to preserve a satisfactory relationship between buyer and supplier.

Sales force—overall image of the manufacturer's sales representatives.

Product quality—perceived quality of a particular product.

The data are found on the text website.

Your task is to complete an exploratory factor analysis on the survey data. The purpose for the consulting group is twofold: (a) to identify the underlying dimensions of these data and (b) to create a new set of variables for inclusion into subsequent assessments of the vendor/distribution channels. Methodology issues to consider in your analysis are:

a Desirability of principal components versus principal axis factoring.

b Decisions on criteria for number of factors to extract.

c Rotation of the factors.

d Factor loading significance.

e Interpretation of the rotated matrix.

Prepare a report summarizing your findings and interpreting your results.

10 A researcher was given the assignment of predicting which of three actions would be taken by the 280 employees in the Desota plant that was going to be sold to its employees. The alternatives were to:

a Take severance pay and leave the company.

b Stay with the new company and give up severance pay.

c Take a transfer to the plant in Chicago.

The researcher gathered data on employee opinions, inspected personnel files and the like, and then did a discriminant analysis. Later, when the results were in, she found the results listed here. How successful was the researcher's analysis?

	Predicted Decision		
Actual Decision	**A**	**B**	**C**
A	80	5	12
B	14	60	14
C	10	15	70

From the Headlines

11 Nokia is the world's leader in cell phones. At the Consumer Electronics Show, Nokia president and CEO, Olli-Pekka Kallasvuo, discussed the company's strategy for reaching the world's developing markets. He wants Nokia to "do good business and do good at the same time." The company's program, "Calling All Innovators," aspires to change people's lives through mobile phones. He described a contest called the "Nokia Growth Economy Venture Challenge" that provides $1 million in funding to a company that create products that improve people's lives. He hopes the venture will promote upward mobility to cell phone users in markets where the average income is less than $5 a day. This year's theme is "Connecting Apps That Make a Difference." What multivariate technique or combination of techniques would you use to investigate mobile applications that, when connected, improve quality of life?

>cases*

Mastering Teacher Leadership

NCRCC: Teeing Up and New Strategic Direction

Proofpoint: Capitalizing on a Reporter's Love of Statistics

* You will find a description of each case in the Case Abstracts section of the textbook. Check the Case Index to determine whether a case provides data, the research instrument, video, or other supplementary material. Written cases are downloadable from the text website (www.mhhe.com/cooper12e). All video material and video cases are available from the Online Learning Center.