>snapshot

What's a Business Education without Wine?

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

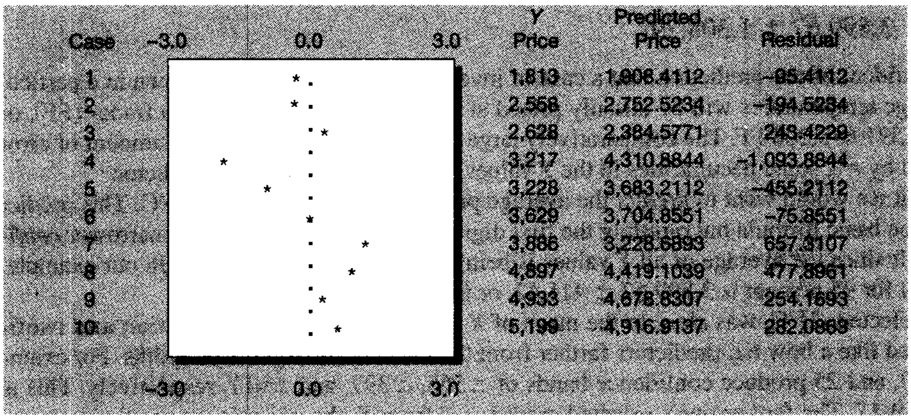
While medical research has shown that moderate drinking can reduce the risk of heart disease, that's not the appeal for students who believe that it can be an effective tool for shaping positive business relationships. Brian Scanlon of Harvard's Wine & Cuisine Society summed it up this way: "Wine knowledge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

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

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

www.winespectator.com

> Exhibit 19-16 Plot of Standardized Residuals: Wine Price Study



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

Comparing the line drawn in Exhibit 19-15 to the trial lines in Exhibit 19-14, one can readily see the success of the least-squares method in minimizing the error of prediction.

Residuals

We now turn our attention to the plot of standardized residuals in Exhibit 19-16. A **residual** is what remains after the line is fit or $(Y_i - \hat{Y}_i)$. When standardized, residuals are comparable to Z scores with a mean of 0 and a standard deviation of 1. In this plot, the standardized residuals should fall between 2 and -2, be randomly distributed about zero, and show no discernible pattern. All these conditions say the model is applied appropriately.

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

Predictions

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

$$\hat{Y} = -645.57 + 216.44(21) = 3{,}899.67$$

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

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

where

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

s = the standard error of estimate (also the square root of the mean square error from the analysis of variance of the regression model) (see Exhibit 19-19).

SS_x = the sum of squares for X (Exhibit 19-13).

$$3{,}899.67 \pm (2.306)(538.559)\sqrt{1 + \frac{1}{10} + \frac{(21 - 19.61)^2}{198.25}}$$

$$3{,}899.67 \pm 1{,}308.29$$

We are 95 percent confident of our prediction that a case of investment-quality red wine grown in a particular year at 21°C average temperatures will be initially priced at 3,899.67 ± 1,308.29 French francs (FF), or from approximately 2,591 to 5,208 FF. The comparatively large band width results from the amount of error in the model (reflected by r^2), some peculiarities in the Y values, and the use of a single predictor.

It is more likely that we would want to predict the average price of *all* cases grown at 21°C. This prediction would use the same basic formula but omitting the first digit (the 1) under the radical. A narrower *confidence* band is the result since the average of all Y values is being predicted from a given X. In our example, the confidence interval for 95 percent is 3,899.67 ± 411.42, or from 3,488 to 4,311 FF.

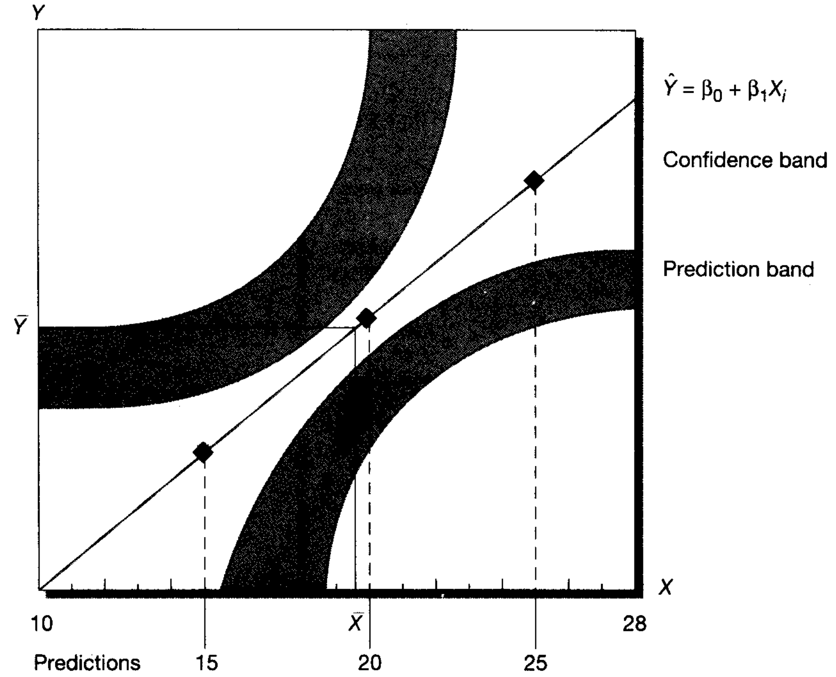
The predictor we selected, 21°C, was close to the mean of X (19.61). Because the **prediction and confidence bands** are shaped like a bow tie, predictors farther from the mean have larger bandwidths. For example, X values of 15, 20, and 25 produce confidence bands of ±565, ±397, and ±617, respectively. This is illustrated in Exhibit 19-17. The farther one's selected predictor is from X, the wider is the prediction interval.

Testing the Goodness of Fit

With the regression line plotted and a few illustrative predictions, we should now gather some evidence of **goodness of fit**—how well the model fits the data. The most important test in bivariate linear regression is whether the slope, β_1, is equal to zero.[11] We have already observed a slope of zero in Exhibit 19-10, line *b*. Zero slopes result from various conditions:

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

> Exhibit 19-17 Prediction and Confidence Bands on Proximity to *X*



The *t*-Test

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

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

where

b_1 was previously defined as the slope β_1.

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

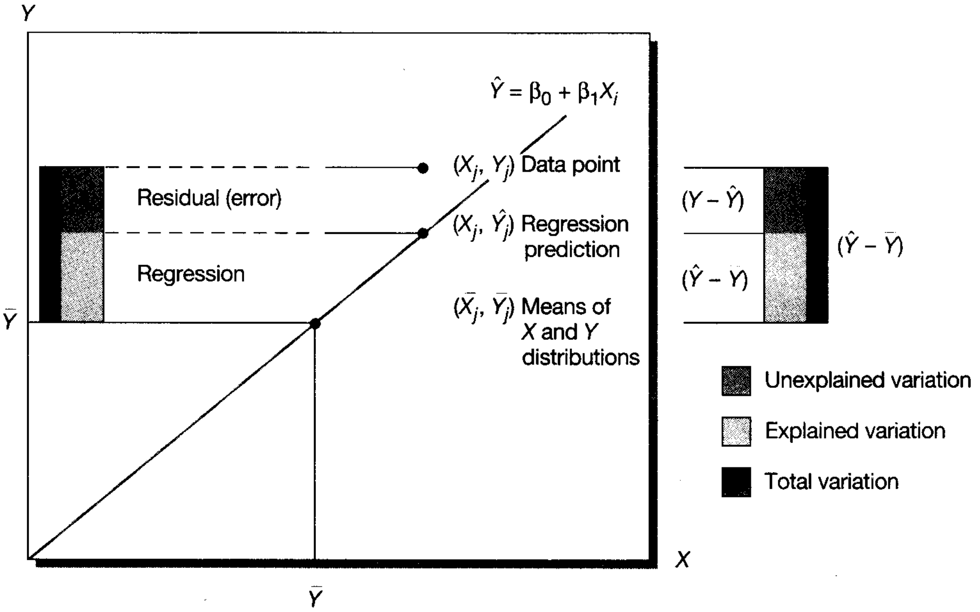
We reject the null hypothesis, $\beta_1 = 0$, because the calculated *t* is greater than any *t* value for 8 degrees of freedom and $\alpha = .01$. Therefore, we conclude that the slope is not equal to zero.

The *F* Test

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

In Exhibit 19-19, we develop this concept sequentially, concluding with the *F* test of the regression model for the wine data. Based on the results presented in that table, we find statistical evidence of a linear relationship between variables. The null hypothesis, $r^2 = 0$, is rejected with $F = 32.02$, d.f. (1, 8), $p < .005$. The

> **Exhibit 19-18** Components of Variation



alternative hypothesis is accepted. The null hypothesis for the F test had the same effect as $\beta_1 = 0$ since we could select either test. Thus, we conclude that X and Y are linearly related.

Coefficient of Determination

In predicting the values of Y without any knowledge of X, our best estimate would be $\bar{Y}$, its mean. Each predicted value that does not fall on Y contributes to an error of estimate, $Y - \bar{Y}$. The total squared error for several predictions would be $\Sigma(Y_i - Y)^2$. By introducing known values of X into a regression equation, we attempt to reduce this error even further. Naturally, this is an improvement over using $\bar{Y}$, and the result is $(\hat{Y} - \bar{Y})$. The total improvement based on several estimates is $\Sigma(\hat{Y}_i - \bar{Y})^2$, the amount of variation explained by the relationship between X and Y in the regression. Based on the formula, the *coefficient of determination* is the ratio of the line of best fit's error over that incurred by using Y. One purpose of testing, then, is to discover whether the regression equation is a more effective predictive device than the mean of the dependent variable.

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

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

> **Exhibit 19-19** Progressive Application of Partitioned Variance Concept

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

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

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

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio
Regression	1	9,287,143.11	9,287,143.11	32.02
Residual (error)	8	2,320,368.49	290,046.06	
Total		11,607,511.60		
				Significance of *F* = .0005

For the wine price study, r^2 was found by using the data from the bottom of Exhibit 19-19:

$$r^2 = 1 - \frac{2,320,368.49}{11,607,511.60} = .80$$

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

> Nonparametric Measures of Association[14]

Measures for Nominal Data

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

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

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

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

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

Chi-Square-Based Measures

< You may wish to review our discussion of chi-square in Chapter 18.

Exhibit 19-20 reports a 2 × 2 table showing the test of an advertising campaign involving 66 people. The variables are success of the campaign and whether direct mail was used. In this example, the observed significance level is less than the testing level ($\alpha = .05$), and the null hypothesis is rejected. A correction to chi-square is provided. We now turn to measures of association to detect the strength of the relationship. Notice that the exhibit also provides an approximate significance of the coefficient based on the chi-square distribution. This is a test of the null hypothesis that no relationship exists between the variables of direct mail and campaign success.

The first **chi-square-based measure** is applied to direct mail and campaign success. It is called **phi (ϕ).** Phi ranges from 0 to +1.0 and attempts to correct χ^2 proportionately to N. Phi is best employed with 2 × 2 tables like Exhibit 19-20 since its coefficient can exceed +1.0 when applied to larger tables. Phi is calculated

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

Phi's coefficient shows a moderate relationship between marketing campaign success and direct mail. There is no suggestion in this interpretation that one variable causes the other, nor is there an indication of the direction of the relationship.

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

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

where k = the lesser number of rows or columns. In Exhibit 19-20, the coefficient is the same as phi.

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

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

> Exhibit 19-20 Chi-Square-Based Measures of Association

	Marketing Campaign Success			
	Count	Yes	No	Row Total
Direct Mail	Yes	21	10	31
	No	13	22	35
	Column Total	34	32	66

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

Minimal expected frequency 15.030

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

*Pearson chi-square probability.

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

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{6.616257}{6.616257 + 66}} =$$ [redacted]

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

Proportional Reduction in Error

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

The computation of lambda is straightforward. In Exhibit 19-21, we have results from an opinion survey with a sample of 400 shareholders in publicly traded firms. Of the 400 shareholders, 180 (45 percent) favor capping executives' salaries; 220 (55 percent) do not favor doing so. With this information alone, if asked to predict the opinions of an individual in the sample, we would achieve the best prediction record by always choosing the modal category. Here it is "do not favor." By doing so, however, we would be wrong 180 out of 400 times. The probability estimate for an incorrect classification is .45, $P(1) = (1 - .55)$.

> **Exhibit 19-21** Proportional Reduction of Error Measures

What is your opinion about capping executives' salaries?

	Cell designation Count Row Pct.	Favor	Do Not Favor	Row Total
Occupational Class	Managerial	1,1 90 82.0	[redacted] 20 18.0	110
	White collar	[redacted] 60 43.0	2,2 80 57.0	140
	Blue collar	[redacted] 30 20.0	3,2 120 80.0	150
	Column Total	[redacted] 45.0%	220 55.0%	[redacted] 100.0%

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

Minimum expected frequency 49.500

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

*Based on chi-square approximation.

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

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

Lambda is then calculated:

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

Note that the asymmetric lambda in Exhibit 19-21, where opinion is the dependent variable, reflects this computation. As a result of knowing the respondents' occupational classification, we improve our prediction by 39 percent. If we wish to predict occupational classification from opinion instead of the opposite, a λ of .24

would be secured. This means that 24 percent of the error in predicting occupational class is eliminated by knowledge of opinion on the executives' salary question. Lambda varies between 0 and 1, corresponding with no ability to eliminate errors to elimination of all errors of prediction.

Goodman and Kruskal's **tau** (τ) uses table marginals to reduce prediction errors. In predicting opinion on executives' salaries without any knowledge of occupational class, we would expect a 50.5 percent correct classification and a 49.5 percent probability of error. These are based on the column marginal percentages in Exhibit 19-21.

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

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

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

When additional knowledge of occupational class is used, information for correct classification of the opinion variable is improved to 62.7 percent with a 37.3 percent probability of error. This is obtained by using the cell counts and marginals for occupational class (refer to Exhibit 19-21), as shown below:

$$\text{Row 1} \quad \left(\frac{90}{110}\right)90 + \left(\frac{20}{110}\right)20 = 73.6364 + 3.6364 = 77.2727$$

$$\text{Row 2} \quad \left(\frac{60}{140}\right)60 + \left(\frac{80}{140}\right)80 = 25.7143 + 45.7142 = 71.4286$$

$$\text{Row 3} \quad \left(\frac{30}{150}\right)30 + \left(\frac{120}{150}\right)120 = 6.0 + 96.0 = \underline{102.0000}$$

Total correct classification (with additional information on occupational class) 250.7013

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

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

Tau is then computed like this:

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

Exhibit 19-21 shows that the information about occupational class has reduced error in predicting opinion to approximately 25 percent. The table also contains information on the test of the null hypothesis that tau = 0 with an approximate observed significance level and asymptotic error (for developing confidence intervals). Based on the small observed significance level, we would conclude that tau is significantly different from a coefficient of 0 and that there is an association between opinion on executives' salaries and occupational class in the population from which the sample was selected. We can also establish the confidence level for the coefficient at the 95 percent level as approximately .25 ± .04.

>snapshot

Speedpass Is McD's Cashless Payment

Lots of people will be waving while visiting McDonald's in Chicago. And it won't be because they want to attract attention or be overly friendly. Instead, some Chicagoland McDonald's (as well as some in Boise, ID) are testing a cashless payment system based on RFID technology by FreedomPay. This system was first tested by McDonald's franchises in New York and Southern California and tested in 2001 by nine Chicago McDonald's restaurant owners. The system, called Speedpass, is activated when a customer waves a Speedpass device at a reader located either in the drive-thru or inside at the checkout counter. Most recently, these devices are designed to dangle from keychains. The Speedpass system was originally introduced by ExxonMobil Corp. at its Mobil gas stations. Similar systems have been tested by Taco Bell and KFC.

How should this study be designed to measure the effectiveness of cashless payment systems? What relationships do you expect to find? Will they require parametric or nonparametric measures of association?

www.speedpass.com; www.mcdonalds.com; www.freedompay.com

Measures for Ordinal Data

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

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

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

Exhibit 19-22 presents data for 70 managerial employees of KeyDesign, a large industrial design firm. All 70 employees have been evaluated for coronary risk by the firm's health insurer. The management levels are ranked, as are the fitness assessments by the physicians. If we were to use a nominal measure of association with these data (such as Cramer's *V*), the computed value of the statistic would be positive since order is not present in nominal data. But using ordinal measures of association reveals the actual nature of the relationship. In this example, all coefficients have negative signs; therefore, lower levels of fitness are associated with higher management levels.

The information in the exhibit has been arranged so that the number of concordant and discordant pairs of individual observations may be calculated. When a subject that ranks higher on one variable also ranks higher on the other variable, the pairs of observations are said to be **concordant.** If a higher ranking on one variable is accompanied by a lower ranking on the other variable, the pairs of observations are **discordant.** Let *P* stand for concordant pairs and *Q* stand for discordant. When concordant pairs exceed discordant pairs in a $P - Q$ relationship, the statistic reports a positive association between the variables under study. As discordant pairs increase over concordant pairs, the association becomes negative. A balance indicates no relationship between the variables. Exhibit 19-23 summarizes the procedure for calculating the summary terms needed in all the statistics we are about to discuss.[15]

Goodman and Kruskal's **gamma** (γ) is a statistic that compares concordant and discordant pairs and then standardizes the outcome by maximizing the value of the denominator. It has a proportional reduction in

> **Exhibit 19-22** Tabled Ranks for Management and Fitness Levels at KeyDesign

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

Statistic	*Value**
Gamma	[redacted]
Kendall's tau *b*	[redacted]
Kendall's tau c	[redacted]
Somers's *d*	
Symmetric	-.51
With fitness dependent	-.53
With management-level dependent	-.50

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

error (PRE) interpretation that connects nicely with what we already know about PRE nominal measures. Gamma is defined as

$$\gamma = \frac{P - Q}{P + Q} = \frac{172 - 992}{172 + 992} = \frac{-820}{1164} =$$

[redacted]

For the fitness data, we conclude that as management level increases, fitness decreases. This is immediately apparent from the larger number of discordant pairs. A more precise explanation for gamma takes its absolute value (ignoring the sign) and relates it to PRE. Hypothetically, if one was trying to predict whether the pairs were concordant or discordant, one might flip a coin and classify the outcome. A better way is to make the prediction based on the preponderance of concordance or discordance; the absolute value of gamma is the proportional reduction in error when prediction is done the second way. For example, you would get a 50 percent hit ratio using the coin. A PRE of .70 improves your hit ratio to 85 percent (.50 × .70) + (.50) = .85.

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

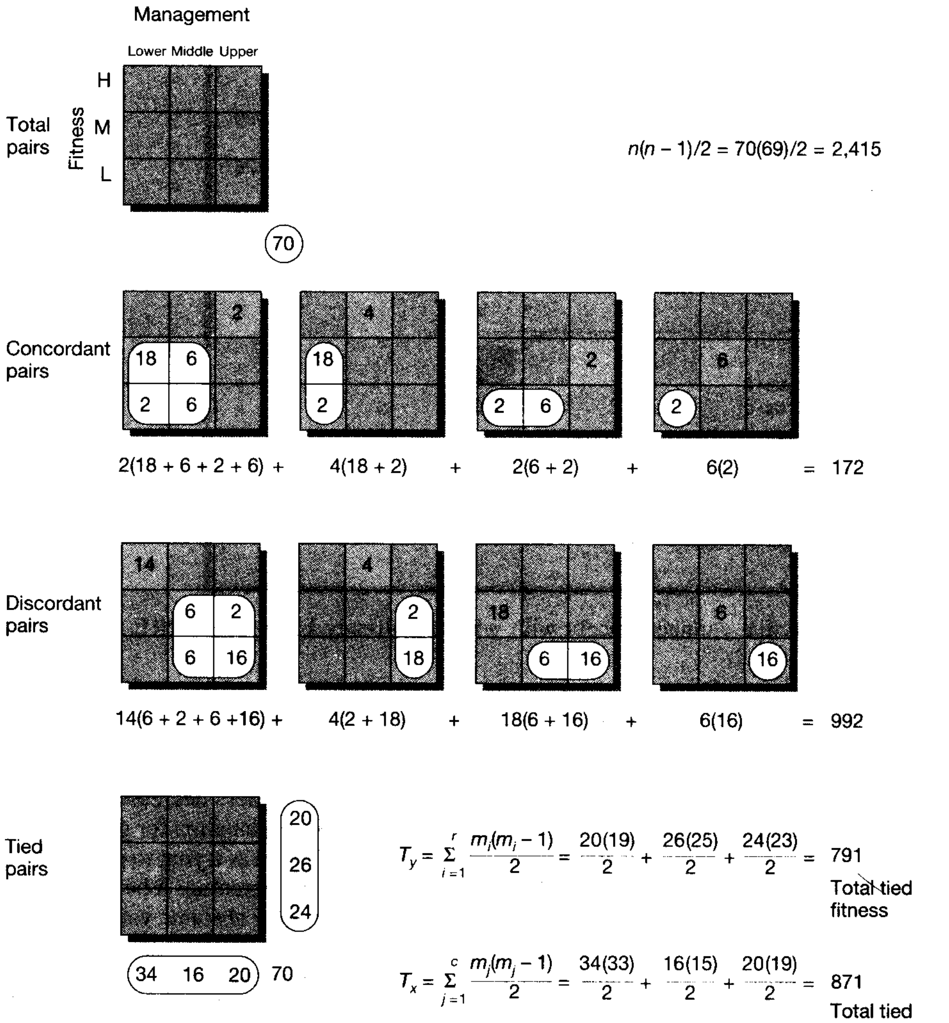
Kendall's **tau *b*** (τ_b) is a refinement of gamma that considers tied pairs. A tied pair occurs when subjects have the same value on the X variable, on the Y variable, or on both. For a given sample size, there are $n(n - 1)/2$ pairs of observations.[18] After concordant pairs and discordant pairs are removed, the remainder are tied. Tau *b* does not have a PRE interpretation but does provide a range of +1.0 to −1.0 for square tables. Its compensation for ties uses the information found in Exhibit 19-23. It may be calculated as

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

$$= \frac{172 - 992}{\sqrt{(2{,}415 - 871)(2{,}415 - 791)}} =$$

[redacted]

> **Exhibit 19-23** Calculation of Concordant (P), Discordant (Q), Tied (T_x, T_y), and Total Paired Observations: KeyDesign Example



$$T_y = \sum_{i=1}^{r} \frac{m_i(m_i - 1)}{2} = \frac{20(19)}{2} + \frac{26(25)}{2} + \frac{24(23)}{2} = 791$$

Total tied fitness

$$T_x = \sum_{j=1}^{c} \frac{m_j(m_j - 1)}{2} = \frac{34(33)}{2} + \frac{16(15)}{2} + \frac{20(19)}{2} = 871$$

Total tied management

where T_x is the total pairs of ties on the column variable
T_y is the total pairs of ties on the row variable
m_{ij} are the marginals

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

$$\tau_c = \frac{2m(P - Q)}{N^2(m - 1)} = \frac{2(3)(172 - 992)}{(70)^2(3 - 1)} =$$ [redacted]

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

Somers's *d* rounds out our coverage of statistics employing the concept of concordant-discordant pairs. This statistic's utility comes from its ability to compensate for tied ranks and adjust for the direction of the dependent variable. Again, we refer to the preliminary calculations provided in Exhibit 19-23 to compute the symmetric and asymmetric *d*'s. As before, the symmetric coefficient (equation 1) takes the row and column variables into account equally. The second and third calculations show fitness as the dependent and management level as the dependent, respectively.

$$d_{\text{sym}} = \frac{(P - Q)}{n(n-1) - T_x T_y/2} = \frac{-820}{1{,}584} = -.51 \qquad (1)$$

$$d_{y-x} = \frac{(P - Q)}{\dfrac{n(n-1)}{2} - T_x} = \frac{-820}{2{,}415 - 871} = -.53 \qquad (2)$$

$$d_{x-y} = \frac{(P - Q)}{\dfrac{n(n-1)}{2} - T_y} = \frac{-820}{2{,}415 - 791} = -.50 \qquad (3)$$

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

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

To illustrate the use of rho, consider a situation where KDL, a media firm, is recruiting account executive trainees. Assume the field has been narrowed to 10 applicants for final evaluation. They arrive at the company headquarters, go through a battery of tests, and are interviewed by a panel of three executives. The test results are evaluated by an industrial psychologist who then ranks the 10 candidates. The executives produce

> **Exhibit 19-24** KDL Data for Spearman's Rho

	Rank by			
Applicant	Panel *x*	Psychologist *y*	*d*	d^2
1	3.5	6.0	−2.5	6.25
2	10.0	5.0	5.0	25.00
3	6.5	8.0	−1.5	2.25
4	2.0	1.5	0.5	0.25
5	1.0	3.0	−2.0	4.00
6	9.0	7.0	2.0	4.00
7	3.5	1.5	2.0	4.00
8	6.5	9.0	−2.5	6.25
9	8.0	10.0	−2.0	4.00
10	5.0	4.0	1.0	1.00
				57.00

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

a composite ranking based on the interviews. Your task is to decide how well these two sets of ranking agree. Exhibit 19-24 contains the data and preliminary calculations. Substituting into the equation, we get

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

where n is the number of subjects being ranked.

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

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

>summary

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

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

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

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

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

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

4 We test regression models for linearity and to discover whether the equation is effective in fitting the data. An important test in bivariate linear regression is whether the slope is equal to zero (i.e., whether the predictor variable *X* is a significant influence on the criterion variable *Y*). In bivariate regression, *t*-tests and *F* tests of the regression produce the same result since t^2 is equal to *F*.

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

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

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

>keyterms

artifact correlations 542

bivariate correlation analysis 535

bivariate normal distribution 539

chi-square-based measures 556

- **contingency coefficient *C*** 556
- **Cramer's *V*** 556
- **phi (ϕ)** 556

coefficient of determination (r^2) 540

concordant 560

discordant 560

error term 548

goodness of fit 552

lambda (λ) 557

linearity 538

method of least squares 548

ordinal measures 560

- **gamma (γ)** 560
- **Somers's *d*** 562
- **Spearman's rho (ρ)** 563
- **tau *b* (τ_b)** 561
- **tau *c* (τ_c)** 562

Pearson correlation coefficient 536

prediction and confidence bands 552

proportional reduction in error (PRE) 557

regression analysis 544

regression coefficients 545

- **intercept (β_0)** 545
- **slope (β_1)** 545

residual 551

scatterplot 537

simple prediction 544

tau (τ) 559

>discussionquestions

Terms in Review

1 Distinguish between the following:

  a Regression coefficient and correlation coefficient.

  b $r = 0$ and $\rho = 0$.

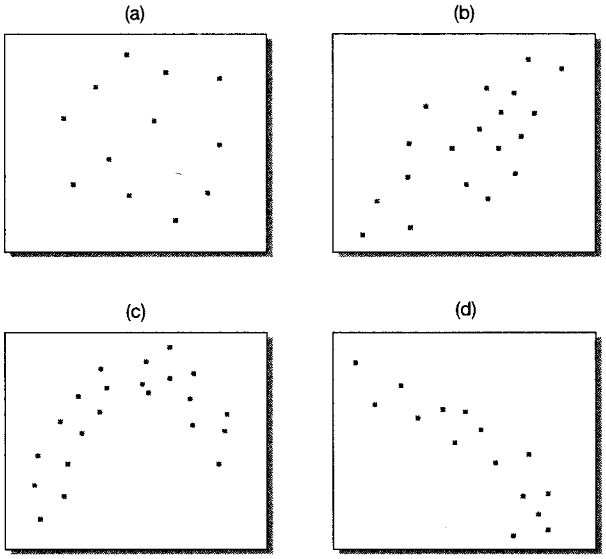
  c The test of the true slope, the test of the intercept, and $r^2 = 0$.

  d r^2 and r.

  e A slope of 0.

  f F and t^2.

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



Making Research Decisions

3 A polling organization collected data on a sample of 60 registered voters regarding a tax on the market value of equity transactions as one remedy for the budget deficit.

Opinion about Market Tax	Education High School	 College Grad.	 MBA
Favorable	15	5	0
Undecided	10	8	2
Unfavorable	0	2	18

a Compute gamma for the table.

b Compute tau *b* or tau *c* for the same data.

c What accounts for the differences?

d Decide which is more suitable for these data.

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

5 A research team conducted a study of soft-drink preferences among residents in a test market prior to an advertising campaign for a new cola product. Of the participants, 130 are teenagers and 130 are adults. The researchers secured the following results:

	Cola	Noncola
Teenagers	50	80
Adults	90	40

Calculate an appropriate measure of association, and decide how to present the results. How might this information affect the advertising strategy?

Bringing Research to Life

6 What would the numbers of "police calls resulting in arrest" for Gladeside and Oceanside need to change to in order to support the conclusion of "disparate impact."

From Concept to Practice

7 Using the following data,

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

a Create a scatterplot.

b Find the least-squares line.

c Plot the line on the diagram.

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

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

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

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

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

Count Actual Clinical Condition	Test Readings of In-Vitro Diagnostic Positive	 Negative	 Total
Pregnant	451 accurate	36 inaccurate	487
Not pregnant	15 inaccurate	183 accurate	198
Total	466	219	685

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

ANOVA Summary Table

	d.f.	Sum of Squares	Mean Square	*F*
Regression	1	11,116,995.47	□	□
Error	□	□	116,104.63	
Total	9	12,045,832.50		

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

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

(See table on next page for data.)

Forbes 500 Random Subsample ($, millions)

Assets	Sales	Market Value	Net Profit	Cash Flow	Number of Employees (thousands)
1,034.00	1,510.00	697.00	82.60	126.50	16.60
956.00	785.00	1,271.00	89.00	191.20	5.00
1,890.00	2,533.00	1,783.00	176.00	267.00	44.00
1,133.00	532.00	752.00	82.30	137.10	2.10
11,682.00	3,790.00	4,149.00	413.50	806.80	11.90
6,080.00	635.00	291.00	18.10	35.20	3.70
31,044.00	3,296.00	2,705.00	337.30	425.50	20.10
5,878.00	3,204.00	2,100.00	145.80	380.00	10.80
1,721.00	981.00	1,573.00	172.60	326.60	1.90
2,135.00	2,268.00	2,634.00	247.20	355.50	21.20

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

11 Using the matrix data (Forbes 500) above, select a pair of variables and run a simple regression. Then investigate the appropriateness of the model for the data using diagnostic tools for evaluating assumptions.

12 For the data below,

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

a Calculate the correlation between X and Y.

b Interpret the sign of the correlation.

c Interpret the square of the correlation.

d Plot the least-squares line.

e Test for a linear relationship:

(1) $\beta_1 = 0$.

(2) $r = 0$.

(3) An F test.

wwwexercise

The University of Michigan's Institute of Social Research is one of the largest education-based survey facilities in the country. Visit its site and read the report on the *National Survey of American Life: Coping with Stress in the 21st Century* **(http://www.rcgd.isr.umich. edu/prba/survey.html).** Click on "questionnaires" and then on two segments of the sample (e.g., "adolescents," "adults"). What's the association of the two groups on any one question.

cases*

Mastering Teacher Leadership

Overdue Bills

NCRCC: Teeing Up and New Strategic Direction

* All cases appear on the text CD; you will find abstracts of these cases in the Case Abstracts section of this text.

>chapter 20

Multivariate Analysis: An Overview

> “Research is formalized curiosity. It is poking and prying with a purpose.”
>
> *Zora Neale Hurston, anthropologist and author*

>**learning**objectives

After reading this chapter, you should understand . . .

1 How to classify and select multivariate techniques.

2 That multiple regression predicts a metric dependent variable from a set of metric independent variables.

3 That 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 The use of cluster analysis techniques for grouping similar objects or people.

9 How perceptions of products or services are revealed numerically and geometrically by multidimensional scaling.

Parker drapes his arm across Sally'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 Sally, drawing Parker's attention as Sally 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 Sally'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 Sally'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. Sally 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

>cont'd

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 Sally, 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 Sally. "I completely forgot to congratulate him on landing the public authority contract and losing the more lucrative one—to us!"

After pausing for effect, Sally 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 where 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.

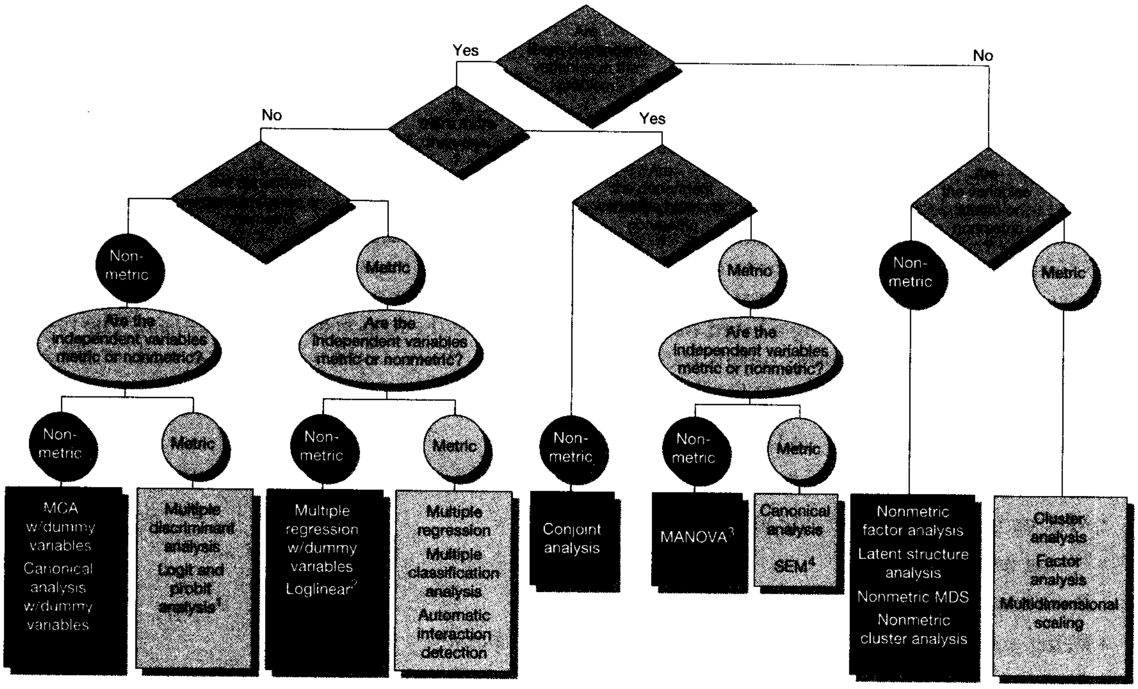
Exhibit 20-1 provides a diagram to guide in the selection of techniques. Let's take an example to show how you might make a decision. Every other year since 1978, the Roper organization has tracked public opinion toward business by providing a list of items that are said to be the responsibility of business. 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 20-1, 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 between 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. *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 20-1, you conclude there is a single dependent variable, family food expenditures. You decide this variable is metric since 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 since it appears to be a dichotomous nominal variable. According to the exhibit, your choices are automatic interaction detection (AID), multiple classification analysis (MCA), and

> Exhibit 20-1 Selecting from the Most Common Multivariate Techniques


Yes
No
No
Yes
Non-metric
Metric
Metric
Non-metric
Metric
Are the independent variables metric or nonmetric?
Are the independent variables metric or nonmetric?
Are the independent variables metric or nonmetric?
Non-metric
Metric
Non-metric
Metric
Non-metric
Non-metric
Metric
MCA w/dummy variables
Canonical analysis w/dummy variables
Multiple discriminant analysis
Logit and probit analysis[1]
Multiple regression w/dummy variables
Loglinear[2]
Multiple regression
Multiple classification analysis
Automatic interaction detection
Conjoint analysis
MANOVA[3]
Canonical analysis
SEM[4]
Nonmetric factor analysis
Latent structure analysis
Nonmetric MDS
Nonmetric cluster analysis
Cluster analysis
Factor analysis
Multidimensional scaling

[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 adapted 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.

multiple regression. You recall from Chapter 17 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.

> 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.[3] 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.[4] 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 19. 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 via a local postal service. 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 some 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 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. 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). Variables may be removed at each step 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 20-2. 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.

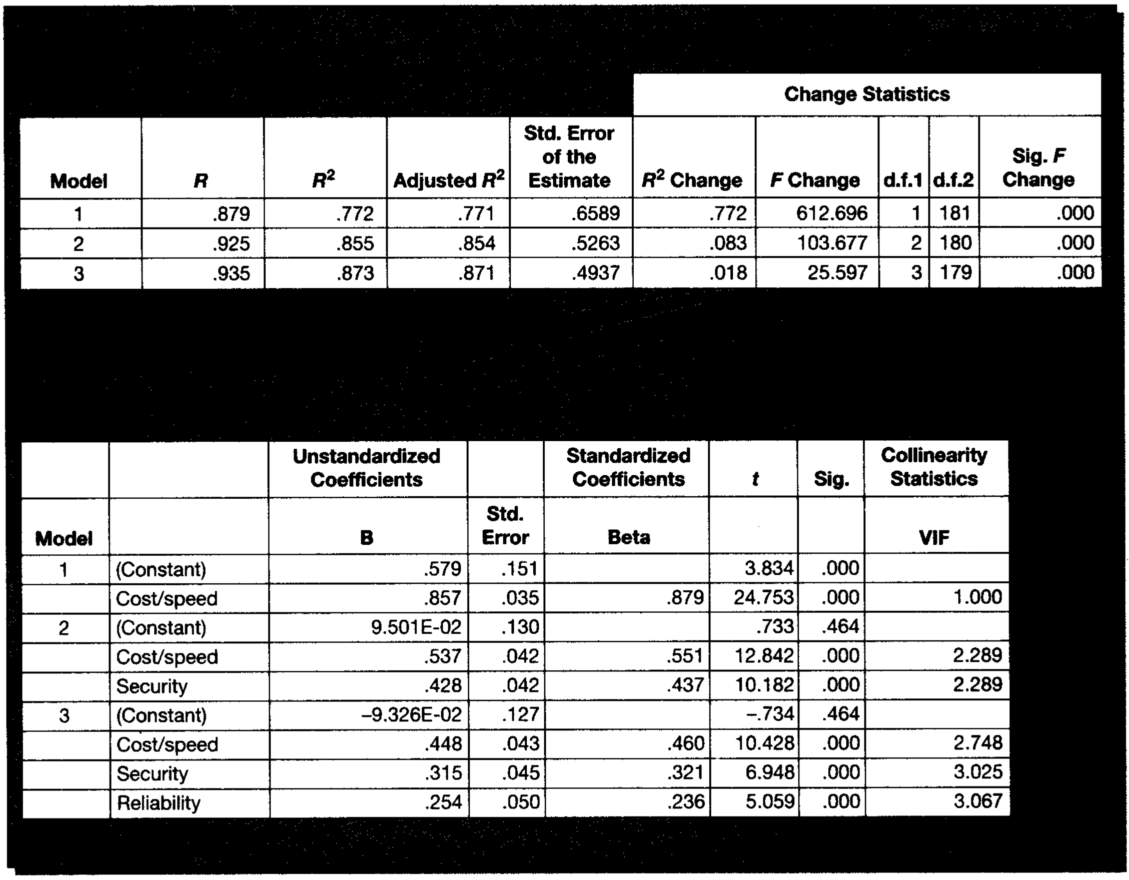
The other reported statistics have the following interpretations.

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^2s 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 C (Exhibit C-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 20-2. 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

> **Exhibit 20-2** Multiple Regression Analysis of Hybrid-Mail Customer Usage, Cost/Speed Valuation, Security, and Reliability



					Change Statistics				
Model	R	R^2	Adjusted R^2	Std. Error of the Estimate	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

		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

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 C, Exhibit C-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,

>snapshot

NCRCC: Teeing Up a New Strategic Direction

NCR Country Club (NCRCC) has undergone a dramatic transformation within the last three years with the construction of a multimillion-dollar clubhouse and dining facility, but the changes have been built on the long-standing tradition of fine golf and dining. Started in 1954 as an employee benefit of the National Cash Register Co. but now an open-membership club, this country club located near Dayton, Ohio, hosts two 18-hole golf courses. The NCRCC South course, a par-71 championship course of 6,824 yards of heavily wooded rolling countryside, has played host to the PGA Championship (1996), the U.S. Open (1986), and the U.S. Mid-Amateur (1998) and is consistently ranked by *Golf Digest* as one of the top 100 courses in the United States.

When its aging membership started to decrease and a one-year membership referral drive didn't dramatically reverse the trend, NCRCC turned to the McMahon Group, a research and strategic golf-course management specialist, for insight and direction. Through an extensive two-stage research design employing six focus groups of 10 to 15 people each, followed by 886 membership surveys, McMahon's research helped NCRCC design a new strategic direction. Sophisticated modeling and analysis led to new facilities for swimming and fitness that turned this proud golf and dining organization into a full-service club for 2,000 with amenities to serve its new target member (the under-46, golf-oriented household, with one or more children under 21 still living at home).

www.mcmahongroup.com; www.ncrcountryclub.com

but correlations at a .80 or greater level should be dealt with in one of two ways: (1) Choose one of the variables and delete the other, or (2) create a new variable that is a composite of the highly intercorrelated variables and use this new variable in place of its components. Making this decision with a correlation matrix alone is not always advisable. In the example just presented, Exhibit 20-2 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. Large values, usually 10.0 or more, suggest collinearity or multicollinearity. With the three predictors in the hybrid-mail example, multicollinearity is not a problem.

< We discuss the correlation matrix, which displays multiple combinations of two variable relationships, in Chapter 19.

Another difficulty with regression occurs when researchers fail to evaluate the equation with data beyond those used originally to calculate it. A practical solution is to set aside a portion of the data (from a fourth to a third) and evaluate the estimating equation. This is called a **holdout sample.** One uses the equation with the holdout data to calculate a new R^2 and compare it to the original R^2 to see how well the equation predicts beyond its data set.

Discriminant Analysis

Researchers often wish to classify people or objects into two or more 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 is to establish a procedure to find the predictors that best classify subjects. Discriminant analysis is frequently used in market segmentation research.

Method

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 found, 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.

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.[5]

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 this. It appears that discriminant analysis is a perfect technique. You begin by gathering data on 30 MBAs who have been hired in recent years. Fifteen of these have been successful employees, while the other 15 have been unsatisfactory. 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

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 20-3. This indicates that 25 of the 30 ($30 - 3 - 2 = 25$) cases have been correctly classified using these three variables.

The standardized and unstandardized discriminant function coefficients are shown in part B of Exhibit 20-3. These results indicate that X_3 (the employment test) has the greatest discriminating power. Several significance tests also may be computed. One, 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 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 dependent variables and classificatory variables or factors. In

> **Exhibit 20-3** Discriminant Analysis Classification Results at KDL Media

A.

			Predicted Success	
Actual Group		Number of Cases	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	.57868
X_3	.53028	.97505
Constant	12.89685	

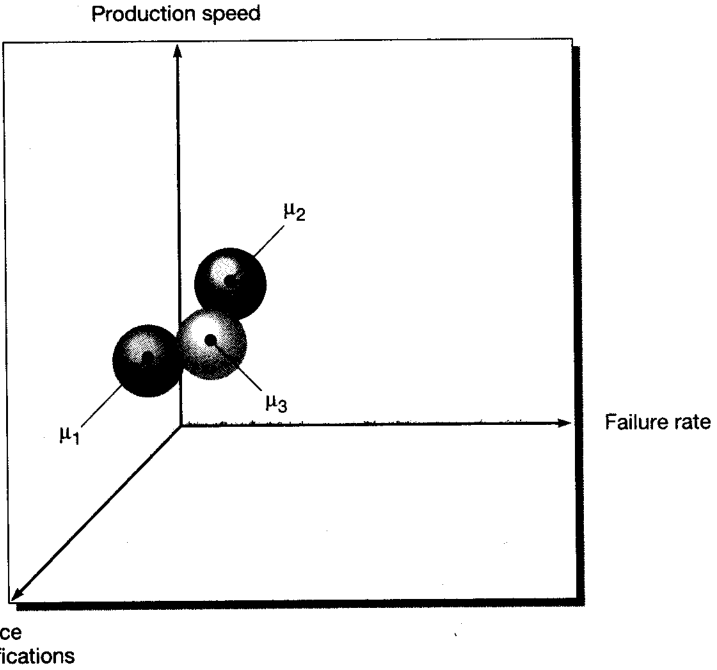
business research, MANOVA can be used to test differences among samples of employees, customers, manufactured products, production parts, and so forth.

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. MANOVA can correct this by simultaneously testing all the variables and their interrelationships. MANOVA uses special matrices [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 treatment 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 20-4 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 alternatives may be considered:

1. Univariate F tests can be run on the dependent variables.
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.[6]

> Exhibit 20-4 MANOVA Techniques Show These Three Centroids to Be Unequal in the CalAudio Study



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.

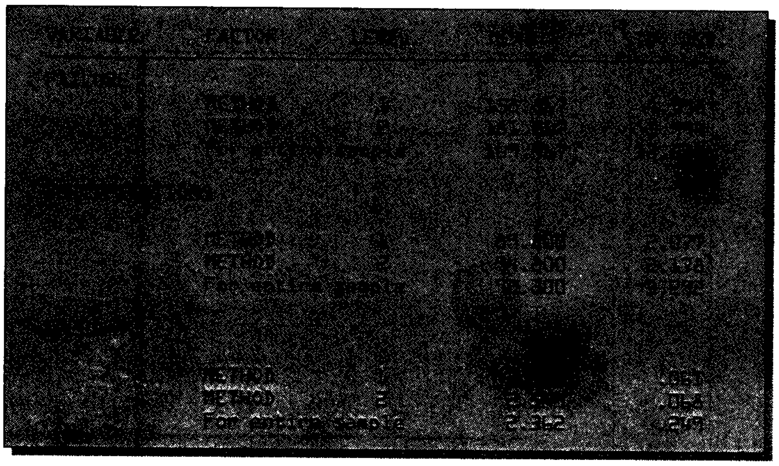
Example

To illustrate, 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

> **Exhibit 20-5** MANOVA Cell Means and Standard Deviations in CalAudio Study



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 20–5 shows the mean and standard deviation of the dependent variables (failure, specifications, and manufacturing speed) for each level of method.[7] 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).

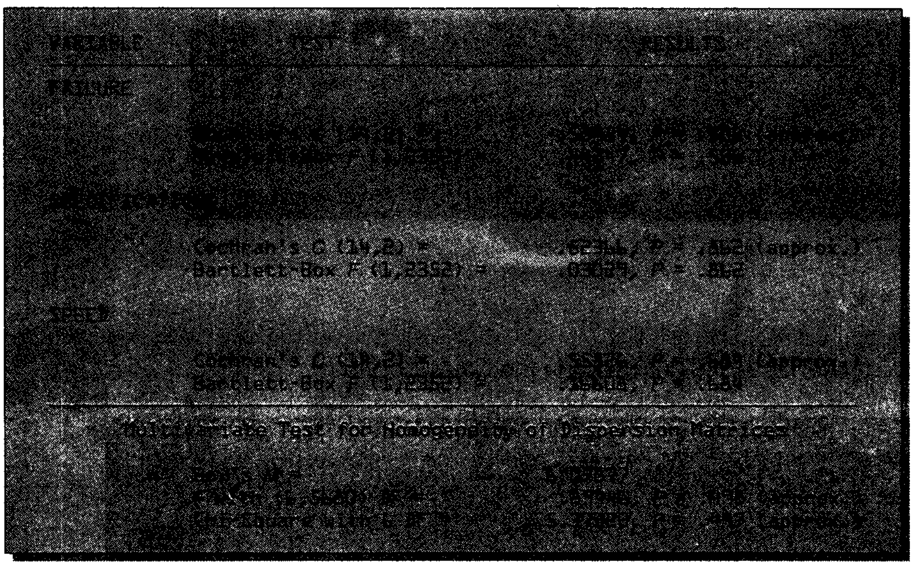
We have used diagnostics to check the assumptions of MANOVA except for equality of variance. Both levels of the manufacturing method variable produce a matrix, and the equality of these two matrices must be determined (H_0: variances are equal). Exhibit 20-6 contains homogeneity-of-variance tests for 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 20-6. Again, we are unable to reject the homogeneity-of-variance assumption regarding the matrices. This satisfies the multivariate assumptions.

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 20-7, 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 20-8 shows three tests, including the Hotelling T^2. All

> **Exhibit 20-6** MANOVA Homogeneity-of-Variance Tests in the CalAudio Study



```
VARIABLE      [illegible] TEST [illegible]           RESULTS

FAILURE       [illegible]

              [illegible]                            [illegible]
              [illegible]                            [illegible]

[illegible]

              Cochran's C (14,2) =                   [illegible], P = [illegible] (approx.)
              Bartlett-Box F (1,2352) =              [illegible], P = [illegible]

SPEED

              Cochran's C (14,2) =                   [illegible], P = [illegible] (approx.)
              Bartlett-Box F (1,2352) =              [illegible], P = [illegible]

   Multivariate Test for Homogeneity of Dispersion Matrices

              Box's M =                              [illegible]
              [illegible] DF =                       [illegible], P = [illegible] (approx.)
              Chi-Square with [illegible] DF =       [illegible], P = [illegible] (approx.)
```

> **Exhibit 20-7** Bartlett's Test of Sphericity in the CalAudio Study

```
Statistics for WITHIN CELLS correlations

[illegible]                                  [illegible]
[illegible]                                  [illegible]

[illegible] criterion =                      [illegible]
```

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 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 20-9. 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 compai isons 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.

< See Chapter 18's discussion of multiple comparison procedures.

Structural Equation Modeling[8]

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.

> **Exhibit 20-8** Multivariate Tests of Significance in the CalAudio Study

[redacted]

Note: F statistics are exact.

> **Exhibit 20-9** Univariate Tests of Significance in the CalAudio Study

[redacted]

Note: F statistics are exact.

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

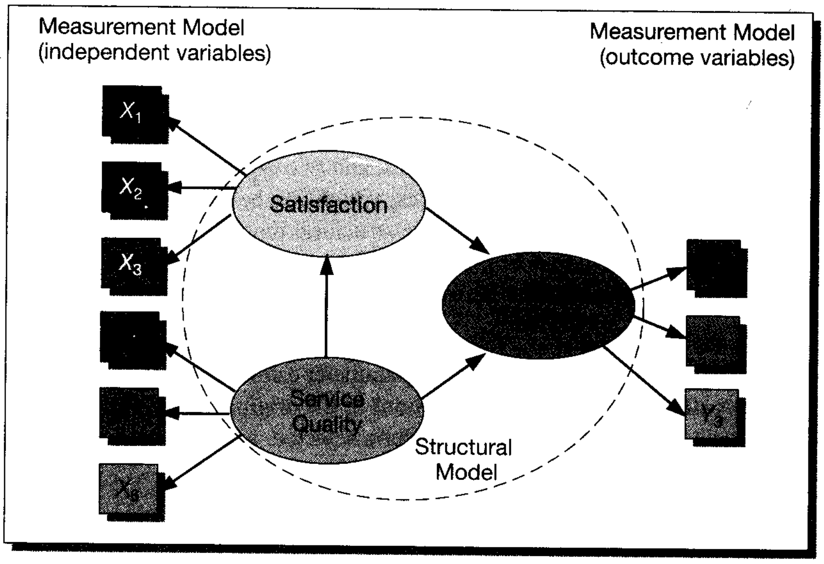
A research consultant, hired by MindWriter, investigated the relationship between customer satisfaction and service quality, as well as the degree to which customer satisfaction and service quality predict customer purchase intention. The researcher used the *competing models strategy,* and proposed 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 20-10).

The example in Exhibit 20-10 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, while 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.

> **Exhibit 20-10** Measurement Models Relative to the Full Structural Equation Model



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 × 3 × 3 × 2 × 2 × 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.

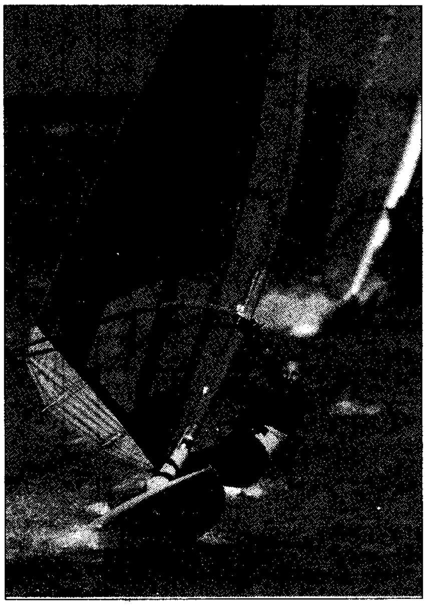
After selecting the factors and their levels, a computer program determines the number of product descriptions necessary to estimate the utilities. SPSS procedures build a file structure for all possible combinations, generate the subset required for testing, produce the card descriptions, and analyze results. The command structure within these procedures provides for holdout sampling, simulations, and other requirements frequently used in commercial applications.[9]

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. In the 1990s, Americans were spending \$1.3 billion on 189 million pairs of sunglasses, most of which failed 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.[10]

Brand	Bolle	Hobbies	Oakley	Ski Optiks
Style*	A B C	A B C	A B C	A B C
[illegible]	[illegible]	[illegible]	Yes No	Yes No
Price	\$100 \$72 \$60 \$40	\$100 \$72 \$60 \$40	\$100 \$72 \$60 \$40	\$100 \$72 \$60 \$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.

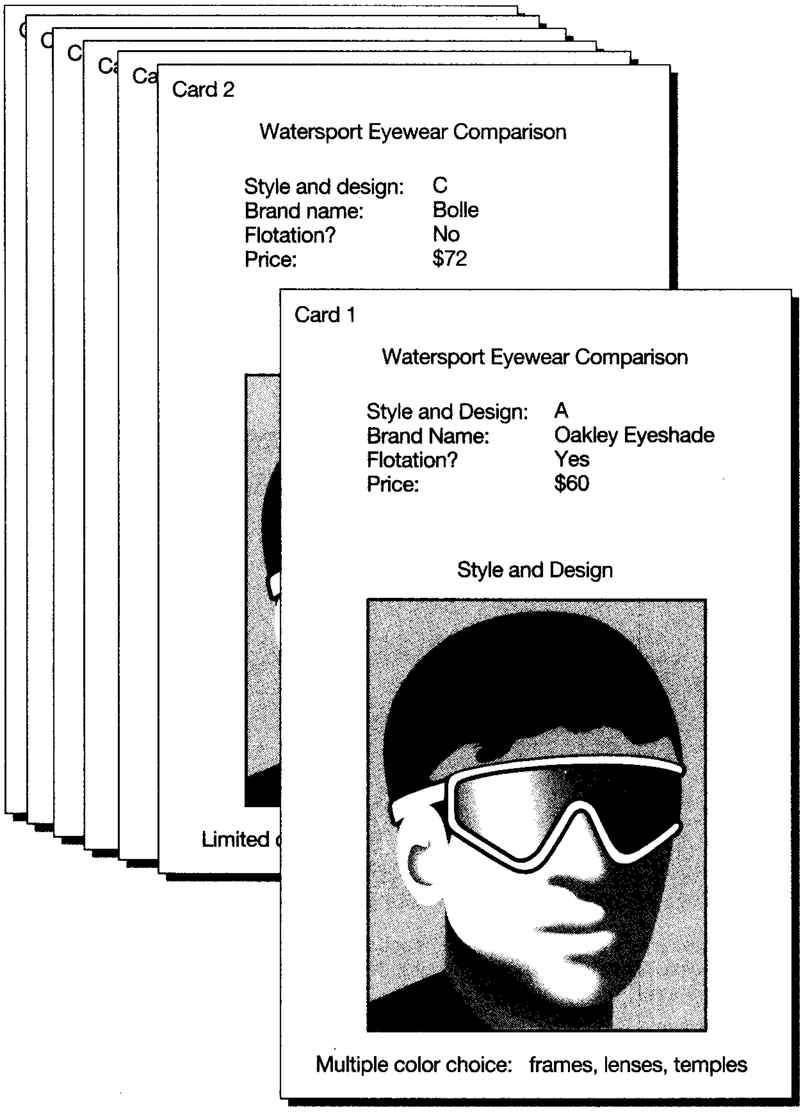


This is a 4 (brand) × 3 (style) × 2 (flotation) × 4 (price) design, or a 96-option full-concept study. The algorithm selected 16 cards to estimate the utilities for the full concept. Combinations of interest that were not selected can be estimated later from the utilities. In addition, four holdout cards were administered to subjects but evaluated separately. The cards shown in Exhibit 20-11 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 20-12 and 20-13.

Exhibit 20-12 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:

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

> Exhibit 20-11 Concept Cards for Conjoint Sunglasses Study



Card 2

Watersport Eyewear Comparison

Style and design:	C
Brand name:	Bolle
Flotation?	No
Price:	$72

Limited

Card 1

Watersport Eyewear Comparison

Style and Design:	A
Brand Name:	Oakley Eyeshade
Flotation?	Yes
Price:	$60

Style and Design

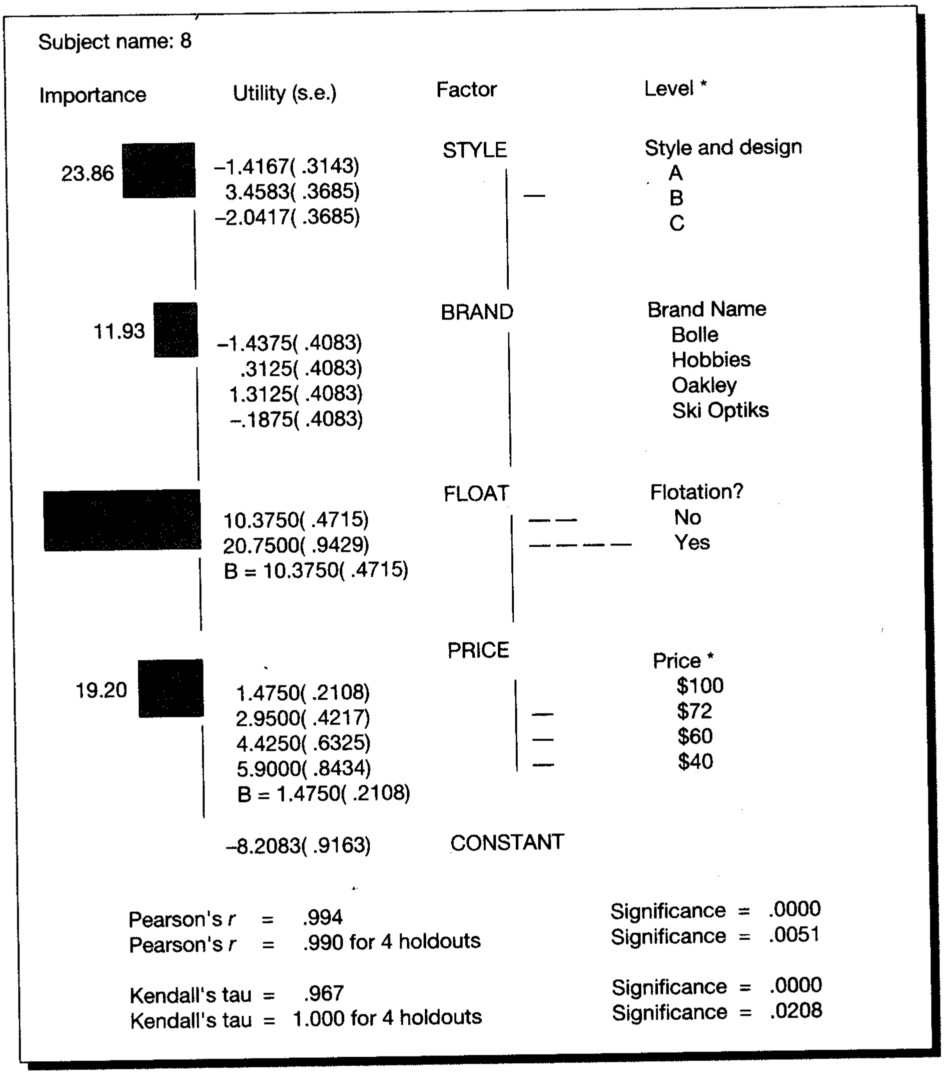
Multiple color choice: frames, lenses, temples

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:

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

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,

> **Exhibit 20-12** Conjoint Results for Participant 8, Sunglasses Study



```
Subject name: 8

Importance      Utility (s.e.)      Factor        Level *

                                    STYLE         Style and design
23.86 ■         -1.4167( .3143)       |             A
                 3.4583( .3685)       |  —          B
                -2.0417( .3685)       |             C
      |                               |
      |                             BRAND         Brand Name
11.93 ■         -1.4375( .4083)       |             Bolle
                  .3125( .4083)       |             Hobbies
                 1.3125( .4083)       |             Oakley
                 -.1875( .4083)       |             Ski Optiks
      |                               |
                                    FLOAT         Flotation?
■■■■■■■■         10.3750( .4715)      | ——          No
                 20.7500( .9429)      | ————        Yes
                 B = 10.3750( .4715)  |
      |                               |
                                    PRICE         Price *
19.20 ■           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
```

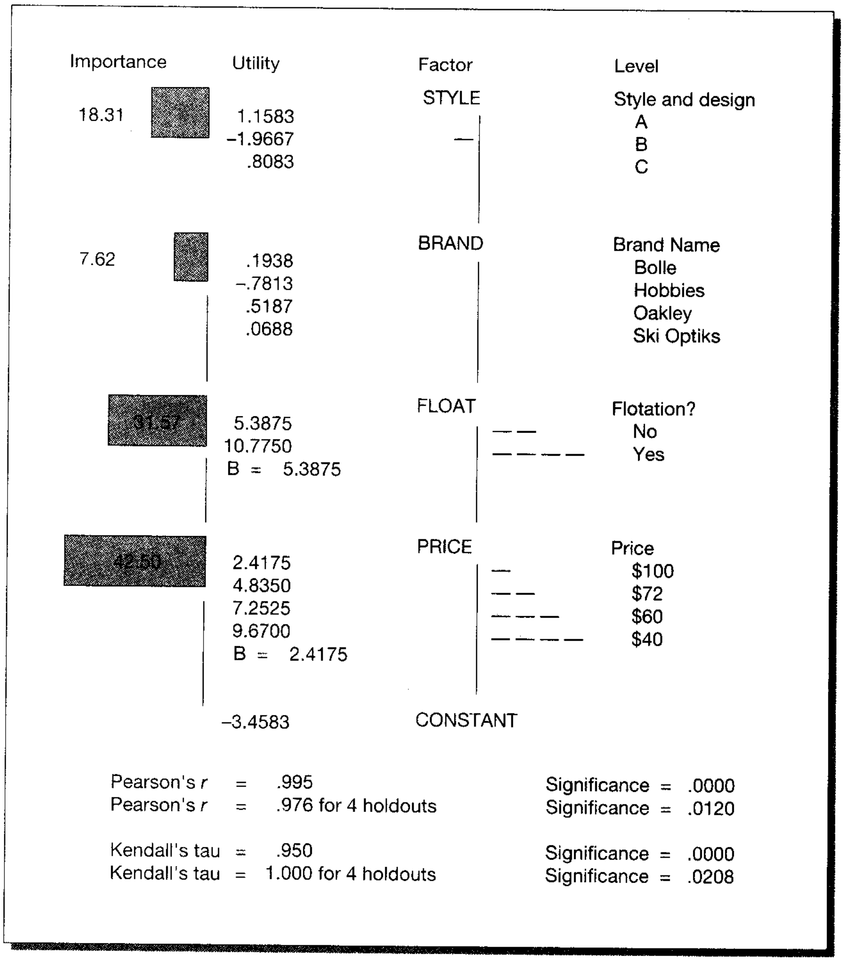
*Subject reversed decision once.

as revealed by the asterisk next to the price factor in Exhibit 20-12. 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 20-13. 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 19. In this application, they measure the relationship between observed and estimated preferences. Since 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.

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 CD for a MindWriter example of conjoint analysis using Simalto+Plus.

> **Exhibit 20-13** Conjoint Results for Sunglasses Study Sample (n = 10)



```
Importance   Utility          Factor       Level

18.31        1.1583           STYLE        Style and design
            -1.9667             —            A
              .8083                          B
                                             C

7.62          .1938           BRAND        Brand Name
             -.7813                          Bolle
              .5187                          Hobbies
              .0688                          Oakley
                                             Ski Optiks

31.57        5.3875           FLOAT        Flotation?
            10.7750                 — —      No
             B =  5.3875            — — — —  Yes

42.50        2.4175           PRICE        Price
             4.8350                 —        $100
             7.2525                 — —      $72
             9.6700                 — — —    $60
             B =  2.4175            — — — —  $40

            -3.4583           CONSTANT

Pearson's r   =  .995                        Significance =  .0000
Pearson's r   =  .976 for 4 holdouts         Significance =  .0120

Kendall's tau =  .950                        Significance =  .0000
Kendall's tau =  1.000 for 4 holdouts        Significance =  .0208
```

> Interdependency Techniques

Factor Analysis

Factor analysis is a general term for several specific computational techniques. All have the objective of reducing to a manageable number many variables that belong together and have overlapping measurement characteristics. The predictor-criterion relationship that was found in the dependence situation is replaced by a matrix of intercorrelations among several variables, none of which is viewed as being dependent on another. For example, one may have data on 100 employees with scores on six attitude scale items.

Method

Factor analysis begins with the construction of a new set of variables based on the relationships in the correlation matrix. While this can be done in a number of ways, the most frequently used approach is **principal components analysis.** This method transforms a set of variables into a new set of composite variables

>snapshot

The Mail as a "Super" E-Business

The world's postal system is projected to grow at a rate of 3.8 percent through 2005, according to its governing body, the Universal Postal Union (UPU). Hybrid mail will account for 6 percent, or 33 billion, of the world's 550 billion pieces of physical mail in 2005 according to the UPU. SuperLetter.com plans to be an e-business success story in this hybrid-mail sector. According to founder and successful entrepreneur Christopher Schultheiss, "We are establishing the world's first global 'hybrid mail' network enabling users to create letters or documents on their personal computers, send them like email in a secure encrypted mode over the Internet to remote printers near the recipients, where they will be printed, folded, enveloped, franked with postage and delivered in the local mail."

Using a variety of multiple-variable analytic techniques, SuperLetter specifically identified its target market as professional and financial service firms, not-for-profit organizations, educational groups, and immigrant/expatriate communities. SuperLetter will also draw from the $100 billion worldwide international courier market, like FedEx, UPS, and DHL, now experiencing strong growth rates (15 percent in international volumes relative to single-digit domestic growth). But the greatest source of messaging is likely to come from the Internet itself. Focused primarily on international correspondence, SuperLetter bridges the gap between conventional door-to-door postal services, which take from 5 to 10 days for overseas delivery, and private express/courier services, which may take from 2 to 3 days. SuperLetter's basic international service delivers a letter from desk to door in 2 to 3 days for about one-tenth of private express costs and under one-half of those costs for same-day services.

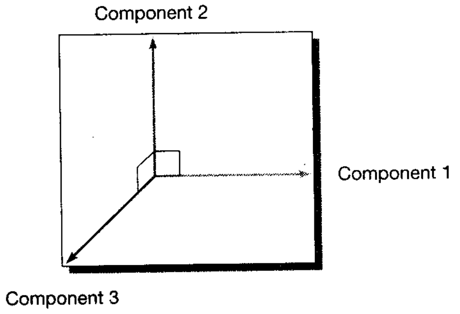
www.superletter.com

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 best combination makes up the first principal component and is the first factor. 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. The output of a principal components analysis might look like the hypothetical data shown in Exhibit 20-14.

Numerical results from a factor study are shown in Exhibit 20-15. 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.** Two other elements in Exhibit 20-15 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

> Exhibit 20-14 Principal Components Analysis from a Three-Variable Data Set



Component no. 1	63%	63%
Component no. 2	29	92
Component no. 3	8	100

> **Exhibit 20-15** 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			

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.

In this case, the unrotated factor loadings are not informative. 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.

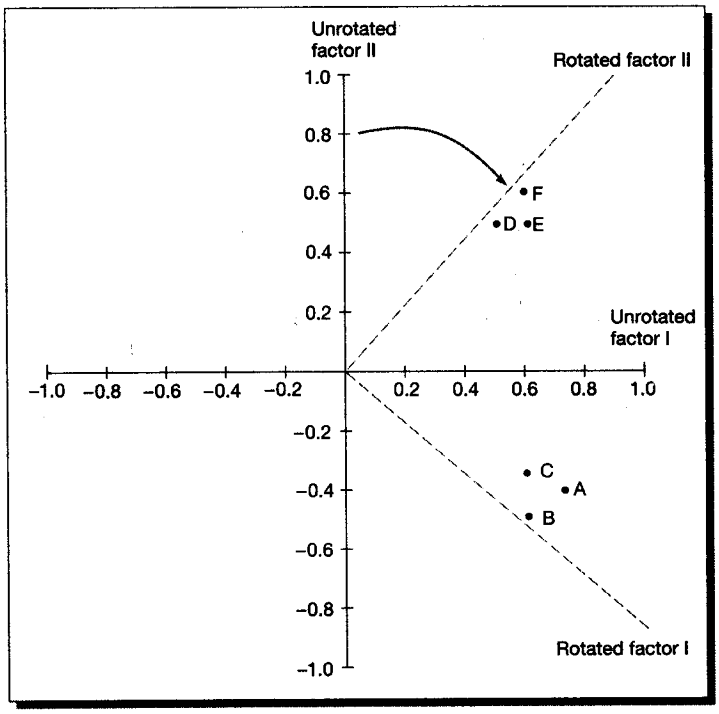
To understand the rotation concept, consider that you are dealing only with simple two-dimensional rather than multidimensional space. The variables in Exhibit 20-15 can be plotted in two-dimensional space as shown in Exhibit 20-16. Two axes divide this space, and the points are positioned relative to 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). From the rotation shown in Exhibit 20-16, it can be seen that the solution is improved substantially. Using the rotated solution suggests that the measurements from six scales may be summarized by two underlying factors (see the rotated factors section of Exhibit 20-15).

The interpretation of factor loadings is largely subjective. There is no way to calculate the meanings of factors; they are what one sees in them. For this reason, factor analysis is largely used for exploration. One can detect patterns in latent variables, discover new concepts, and reduce data. Factor analysis is also applied to test hypotheses with confirmatory models using SEM.

Example

Student grades make an interesting example. The chairperson of Metro U's MBA program has been review-

>**Exhibit 20-16** Orthogonal Factor Rotations



ing 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 a sample of 21 grade reports is chosen for students in the middle of the GPA range. Three steps are followed:

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 20-17 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 correlation matrix, the extraction of components is shown in Exhibit 20-18. While the program will produce a table with as many as 10 factors, you choose, in this case, to stop 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 20-18 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:

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

> Exhibit 20-17 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 20-18 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	0.26	−.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 percent of variance accounted for by each factor in Exhibit 20-18 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.

In an effort to further clarify the factors, a varimax (orthogonal) rotation is used to secure the matrix shown in Exhibit 20-19. 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				

> **Exhibit 20-19** 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

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.

A number of problems affect the interpretation of these results. Among the major ones are these:

1. The sample is small and any attempt at replication might produce a different pattern of factor loadings.
2. From the same data, another number of factors rather than three can result in different patterns.
3. Even if the findings are replicated, the differences may be due to the varying influence of professors or the way they teach the courses rather than to the subject content.
4. The labels may not truly reflect the latent construct that underlies any factors we extract.

This suggests that factor analysis can be a demanding tool to use. It is powerful, but the results must be interpreted with great care.

Cluster Analysis

Unlike techniques for analyzing the relationships between variables, **cluster analysis** is a set of 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 classifying, or "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 where the goal is to classify similar groups. It shares some similarities with factor analysis, especially when factor analysis is applied to people (*Q*-analysis) instead of to variables. 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. 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.

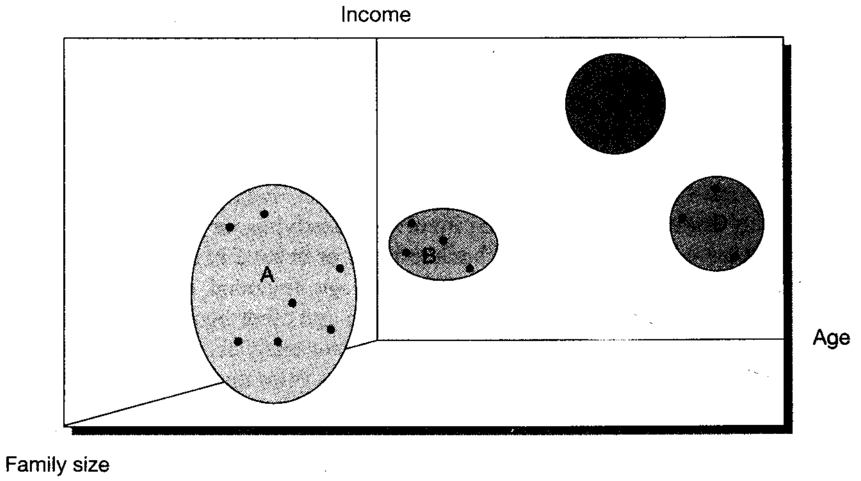
Different clustering methods can and do produce different solutions. It is important to have enough information about the data to know when the derived groups are real and not merely imposed on the data by the method.

The example in Exhibit 20-20 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 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

> **Exhibit 20-20** Cluster Analysis on Three Dimensions



> **Exhibit 20-21** 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

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.

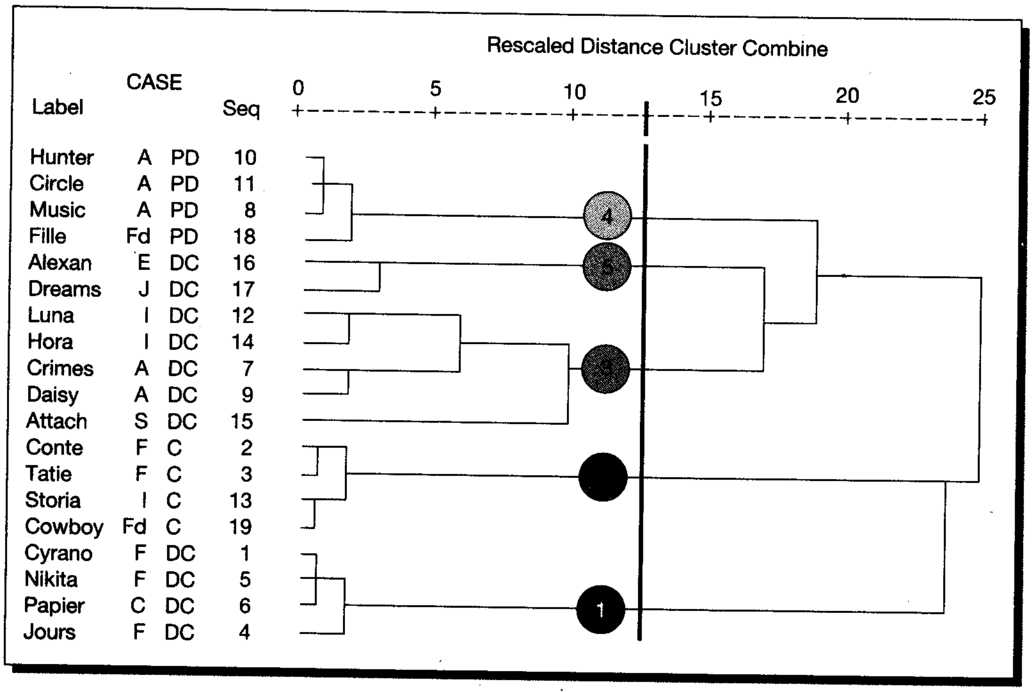
We selected ratings from 12 cinema reviewers using sources ranging from *Le Monde* to international publications sold in Paris. The reviews reputedly influence box-office receipts, and the entertainment business takes them seriously.

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 20-21 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 20-22 shows that the clustering procedure begins with 19 films and continues until all the films are again an undifferentiated group. The solid vertical line shows the point 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 20-22 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

> **Exhibit 20-22** Dendogram of Film Study Using Average Linkage Method



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).

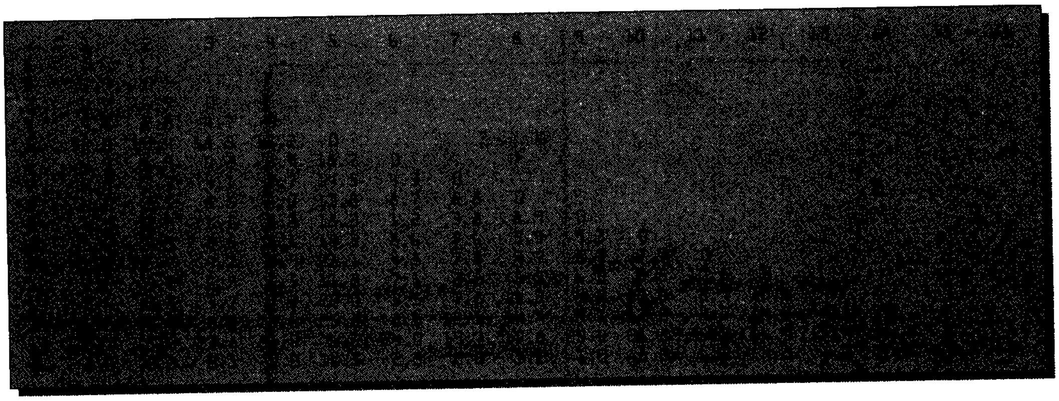
Multidimensional Scaling

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.

> **Exhibit 20-23** Similarities Matrix of 16 Restaurants



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.[11] 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 20-23. 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.

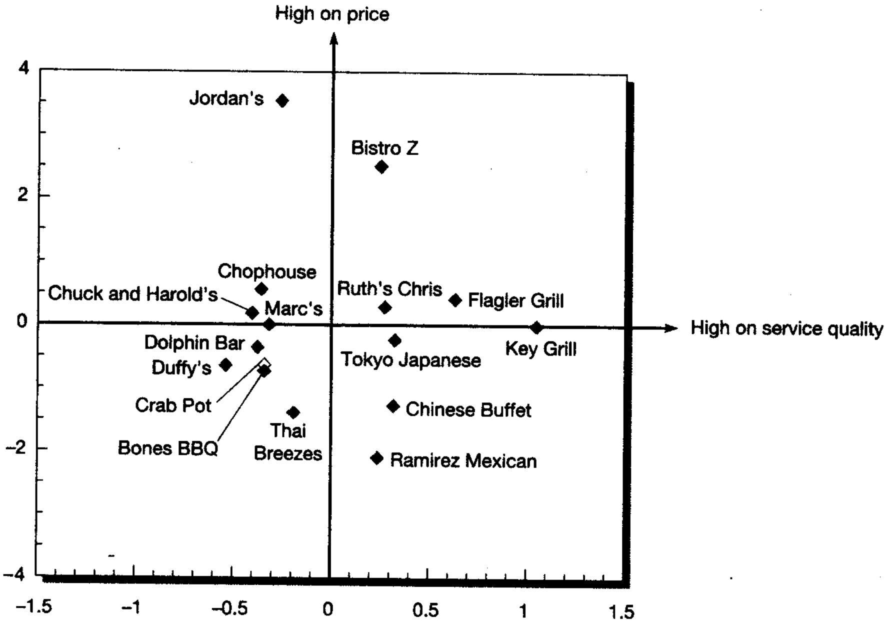
A computer program is used to analyze the data matrix and generate a perceptual map.[12] The objective is to find a multidimensional spatial pattern that best reproduces the original order of the data. For example, the most similar pair (restaurants 3, 6) must be located in this multidimensional space 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 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 20-24. The distance between Crab Pot and Bones BBQ (3, 6) is the shortest, while

> Exhibit 20-24 Positioning of Selected Restaurants



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.

>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.