A walk through discriminant analysis

By Michael Lieberman

Editor's note: Michael Lieberman is founder and president of Multivariate Solutions, a New York consulting firm. He can be reached at 212-656-1711 or at michael@mvsolution.com.

Recently, after conducting a successful market segmentation for a client (we were able to identify high-likelihood customers who are price-insensitive), my client phoned me. "Michael," he said, "my client loves the segments. He wants to be able to run that banner point in the next study. Is there a way to add a few questions to the survey and come up with the classifications?"

"Yes," I answered. "What you need is a discriminant analysis."

At first glance this is not what my client requested. He wants to identify people, not classify them. What, then, was he asking for?

What my client's client wanted to do, in essence, was to discriminate between segment members and non-segment members. Once identified, segment members will be used in a future banner for analysis in the next study.

Discriminant analysis is used in situations where you want to build a predictive model of group membership based on observed data — characteristics, attitudes, demographic attributes, etc. The analysis produces a linear equation of variables that can be used to explain which attribute best discriminates between the two groups and, as an extension, build a powerful predictive model for future classification.

Sometimes clients confuse discriminant analysis with cluster analysis. In fact, they are conceptually similar. However, one uses cluster analysis to form groups. Discriminant flows in the opposite direction: You have the groups, you want to know why.

Basics of discriminant analysis

Discriminant analysis is an *a priori* technique. That is, you have the groups defined before you begin. Multiple discriminant analysis, from which discriminant maps are drawn, is a case where you have membership from more than one group. For ease of understanding, we are going to restrict our case to a simple discriminant with definition of two groups.

Characteristics of the grouping variable are simple. They are distinct, mutually exclusive, and exhaustive. In the case of my client's request, either a respondent is in the target group or he isn't. No fence-sitting. No overlapping.

Basic data assumptions of the predictor variables are that they are normally distributed and independent.

Choosing which predictor variables will be included in the analysis requires a bit of marketing sense. For example, our client seeks to distinguish between high-probability customers and low-probability customers. Within the survey, respondents are asked to rate the company on a given array of attributes — rankings of importance, performance, company image, and firm demographics such as size, revenue, number of employees, and geographic area. A good analysis, especially if it is going to be used for back classification, cannot use all the data available. The results would be murky and there would be a good deal of variation error, commonly referred to as noise.

Therefore, it is vital to choose which predictors go into the equation. In our fictitious example, similar to the case above, attitudes toward the technical prowess of the firm, marketing support, customer service, size of firm, and revenues were chosen.
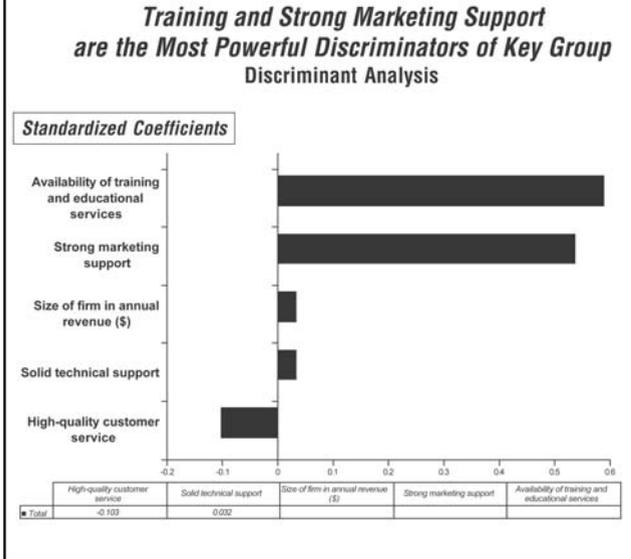
The output

The analysis produces a discriminant function. That is, a linear equation where coefficients are multiplied against the values of the predictor variables to produce a discriminant score. Derived from the discriminant score, a likelihood of each group membership is calculated based on past group membership. To put it simply, the respondent fills out the form and gets a score, which is then compared to a chart to see if he qualifies for the group.

As in all sophisticated statistical analysis, a blizzard of output accompanies the procedure. There are five outcomes that I examine and report: the beta scores of the discrimi-

Variables Present in Model Equation	Standardized Canonical Discriminant Function Coefficients
Availability of training and educational services	0.590
Strong marketing support	0.537
Solid technical suppor	0.032
Size of firm in annual revenue (S)	0.032
Highly quality customer service	-0.103

Figure 2

Training and Strong Marketing Support are the Most Powerful Discriminators of Key Group
Discriminant Analysis

Standardized Coefficients

	High-quality customer service	Solid technical support	Size of firm in annual revenue ($)	Strong marketing support	Availability of training and educational services
■ Total	-0.103	0.032			

For interpreting the standardized coefficients it is more useful to look at them relative to each other. "Availability of training and educational services" has a coefficient of .590. The next attribute, "Strong marketing support" has a coefficient of .537. What this means is that these two attributes are the strongest indicators of membership to the group. "Solid technical support" (.032), "Size of firm in annual revenue" (.032), and "High-quality customer service" (-.103), are near zero and, thus, not strong indicators.

The strengths of the standardized coefficients are relative to each other. A rule is that if a predictor has twice the standardized coefficient of another predictor, it is twice as good a discriminator for the group. Predictors near zero have lit-

Figure 3

Variables Present in Model Equation	Raw Canonical Discriminant Function Coefficients
Availability of training and educational services	1.160
Strong marketing support	0.902
Size of firm in annual revenue (S)	0.072
Solid technical suppor	0.067
Highly quality customer service	-0.273

tle effect. Figure 2 graphically displays these results.

The marketing interpretation for this model is clear. Technical support, size of the company, and customer service are not a major concern to customers when approaching potential suppliers. Though determined conclusively through regression analysis, another conclusion is that if the respondent believes the company has good training and marketing support, he is probably a good candidate to become a customer.

Figure 3 displays raw coefficients. These can be descriptive too, though I tend to use the standardized coefficients for pure descriptive purposes (due to differing scales among the predictor variables; "standardized" gives each predictor a mean of 0 and a standard deviation of 1). Unlike standardized, raw coefficients serve a dual function. They are also used in the re-classification phase.

Finally, it is useful to assess the model itself. Shown also in Figure 4 is the Wilk's Lambda and the percentage reclassified correctly. The Wilk's Lambda is clearly significant.

nant function (known as the raw coefficients), the standardized coefficients, Wilk's Lambda, the discriminant score, and the percentage of respondents correctly reclassified based on the function once it is rerun.

The raw and standardized coefficients are used for descriptive and classification purposes. The discriminant score, when calculated afterwards, is the instrument used for future classification. Wilk's Lambda is a statistic that gives us the robustness of the model. Wilk's Lambda includes a chi-square test, which, if significant, says that the model has tested well and can be assumed strong and reasonably accurate. The percentage of correctly classified respondents tells us how many people returned to where they belong once rerun through the model. It, like Wilk's Lambda, is a measure of how good the model is.

The analysis — descriptive aspects

The groups are defined (potential customers and non-potential customers), the predictor variables are chosen and properly recoded and the analysis is run. My firm uses SPSS to perform the function. This, as I mentioned, produces a large amount of output. To make the outcome simple and actionable, we often transfer them to an Excel spreadsheet or PowerPoint slides which are easy for our clients to understand and incorporate into their reports.

Figure 1 shows the five parameters of our fictitious example, ranked in descending order. Figure 2 is a graphic display.

From here I will walk through the example in the same order as if I were delivering it to a client. The first output I would report, for descriptive purposes, is the standardized coefficients.

Figure 4

Wilks' Lambda	Chi-square	Sig.
0.241	133.023	0.000

67.4% of original grouped cases correctly classified

Look at the "Sig." column, which reads something like "This is the chance the model is zero, or meaningless." In our example the Sig. is 0.000, or 0 percent chance. Generally I accept any model with a Wilk's Lambda Sig. less than 10 percent. With more than two-thirds, 67.4 percent, of respondents being correctly reclassified, we can be confident that model is robust and the process a good fit.

The analysis — predictive aspects

Great. We have the discriminators. The client now wants to be able to reclassify future studies according to the groups already in existence. In addition, for a purely promotional application, the client wants to be able to phone a potential customer, ask him a few questions, and determine if he is a

good candidate for a follow-up.

First fact: In order to successfully perform a re-classification, you must ask exactly the same questions that are present in the model. Also, you must use the same scales.

The process is as follows: Ask the questions and plug the answers back into the equation. The model will produce a discriminant score. The prior run has produced a look-up table of sorts which shows discriminant scores and the likelihood of a person with that score joining the group. In practice, if a given respondent has a score with a corresponding likelihood higher than 50 percent, put him in the group.

For a large number of respondents, my firm will write a small SPSS syntax program so that the process of re-classification for a large number of data will become automated. That is, the banner points can be re-created from the previous study.

For individual respondents an Excel spreadsheet calculator is built. Shown in Figure 5, it is used to calculate the discriminant score and then compare the derived value to the values that are presented in the look-up table. This is used for a client who wishes, say, to phone a prospective customer, ask him a few questions, and then decide if he has a reasonable chance of becoming a real customer.

In our example, the respondent has a score of 7.774. We go to Figure 6 — output from SPSS which gives calculated, existing discriminant scores and the probability a respondent with that score will end up in the group — and find that this score corresponds with a likelihood between 67 percent and 72 percent to belong to our target group. Conclusion: call him back.

Useful and popular

Discriminant analysis is one a number of statistical tech-

niques that we offer to our clients in order to add value to existing projects or pre-plan for a larger data delivery within the context of expected output. The power and efficiency of the process allows strategic planners to capitalize on the existing data to explain and predict consumer conduct without consulting a mystic. It is among our more useful and popular techniques given its power and ease of use.

Figure 5

	Discriminant Coefficents	Attribute	Score Contribution	Key For Rating Variables
Availability of training and educational services	1.160	4	4.640	5=Extremely Satisfied
Strong marketing support	0.902	4	3.608	4=Very Satisfied
Size of firm in annual revenue ($)	0.072	2	0.144	3=Somewhat Satisfied
Solid technical support	0.067	3	0.201	2=Very Unsatisfied
Highly quality customer service	-0.273	3	-0.819	1=Extemely Unsatisfied
				0=Don't know
Discriminant Score			7.774	

Annual revenues ($)
1=$1,000,000 or Under
2=$1,000,000 to $5,000,000
3=$5,000,000 or More

Figure 6

Discriminant Score (From Calculation Table)	Probabilities of Membership in Customer Group
10.45	100.0%
9.20	100.0%
8.40	99.0%
8.35	98.0%
8.24	92.8%
8.12	87.6%
8.01	82.4%
7.90	77.2%
7.79	72.0%
7.67	66.8%
7.56	61.6%
7.45	56.4%
7.33	51.2%
7.22	46.0%
7.11	40.8%
6.99	35.6%
6.88	30.4%
6.77	25.2%
6.66	20.0%
6.54	14.8%
6.43	9.6%
6.31	5.0%
6.26	4.0%
5.03	0.0%
4.60	0.0%
4.55	0.0%
4.22	0.0%
4.17	0.0%
3.37	0.0%
2.09	0.0%
1.93	0.0%