ABPDEA
Associação Brasileira para a Proteção dos Direitos Editoriais e Autorais
RESPEITE O AUTOR
NÃO FAÇA CÓPIA
www.abpdea.org.br

M249p Malhotra, Naresh
Pesquisa de marketing : uma orientação aplicada / Naresh Malhotra ; tradução Laura Bocco. – 4. ed. – Porto Alegre : Bookman, 2006.
720 p. ; 28 cm.

ISBN 85-363-0650-5

1. Marketing – Pesquisa. I. Título.

CDU 658.8.012.12

Catalogação na publicação: Júlia Angst Coelho – CRB Provisório 05/05

Naresh K. Malhotra
Georgia Institute of Technology

PESQUISA DE MARKETING
UMA ORIENTAÇÃO APLICADA

4ª EDIÇÃO

Tradução:
Laura Bocco

Consultoria, supervisão e revisão técnica desta edição:
Tânia Modesto Veludo-de-Oliveira
Mestre em Administração pela FEA/USP

Bookman
2006

CAPÍTULO 18

Análise Discriminante

"Muitas vezes, você mediu diferentes grupos de entrevistados com muitas variáveis métricas. A análise discriminante constitui uma maneira útil de responder as perguntas: Os grupos são diferentes? Em que variáveis são mais diferentes? Posso prever a que grupo uma pessoa pertence usando essas variáveis?"

Jamie Baker-Prewitt, vice-presidente, consultoria e análise, Burke, Inc.

OBJETIVOS

Após a leitura deste capítulo, o aluno deverá estar apto a:

1. Definir o conceito de análise discriminante, seus objetivos e suas aplicações na pesquisa de marketing.
2. Delinear os procedimentos para a realização da análise discriminante, entre eles a formulação do problema, a estimação dos coeficientes da função discriminante, a determinação da significância, a interpretação e a validação.
3. Discutir a análise discriminante múltipla e a distinção entre análise discriminante de dois grupos e múltipla.
4. Explicar a análise discriminante passo a passo e descrever o procedimento Mahalanobis.

Aspectos gerais

Este capítulo aborda a técnica da análise discriminante. Começamos pelo exame da relação deste procedimento com a análise de regressão (Capítulo 17) e a análise da variância (Capítulo 16). Apresentamos um modelo e detalhamos um procedimento geral para fazer análise discriminante, com ênfase na formulação, estimação, determinação da significância, interpretação e validação dos resultados. O procedimento é ilustrado com um exemplo de análise discriminante de dois grupos, seguido por um exemplo de análise discriminante múltipla (de três grupos). Inclui-se também o procedimento de análise discriminante passo a passo.

PESQUISA ATIVA

PROJETO DA LOJA DE DEPARTAMENTOS

Análise discriminante de dois grupos

No projeto da loja de departamentos, utilizou-se a análise discriminante de dois grupos para determinar se os entrevistados mais familiarizados com as lojas, em contraste com os não-familiarizados, atribuíam importâncias relativas diferentes aos oito fatores dos critérios de escolha. Os dois grupos constituíram a variável dependente de familiaridade, e a variável independente foi a importância atribuída aos oito fatores dos critérios de escolha. A função discriminante se revelou significativa, indicando diferenças importantes entre os dois grupos. Os resultados mostraram que, em comparação com os não-familiarizados, os entrevistados familiarizados atribuíam maior importância relativa à qualidade dos produtos, às normas de devolução e de ajuste, ao atendimento por parte dos empregados das lojas e à política de crédito e faturamento.

PESQUISA REAL

Defensores do desconto

Foi feito um estudo com 294 consumidores para determinar a propensão para o desconto, ou as características de consumidores que reagem favoravelmente a promoções com desconto. As variáveis previsoras foram quatro fatores relacionados com as atitudes e comportamentos em relação a compras e características demográficas selecionadas (gênero, idade e renda). A variável dependente foi o grau de propensão do entrevistado em relação ao desconto, tendo-se identificado três níveis dessa propensão. Os entrevistados que não fizeram nenhuma compra motivada pelo desconto nos últimos 12 meses foram classificados como não-usuários; os que fizeram uma ou duas dessas compras foram enquadrados como usuários eventuais, e os que apresentaram mais de duas compras, como usuários freqüentes das compras com desconto. Os dados foram estudados com auxílio da análise discriminante múltipla.

Surgiram logo dois resultados importantes. Em primeiro lugar, a percepção dos consumidores da relação esforço/valor foi a variável mais efetiva na discriminação entre usuários freqüentes, eventuais e não-usuários das ofertas com desconto. É claro que os consumidores favoráveis ao desconto associam menor esforço ao preenchimento das formalidades das compras com desconto e estão dispostos a aceitar um reembolso relativamente menor do que outros consumidores. Em segundo lugar, os consumidores conscientes dos preços normais dos produtos e que podem reconhecer um bom negócio têm probabilidade muito maior de reagir favoravelmente a ofertas com desconto.

Essas constatações foram utilizadas pela DIRECTV, Inc. (*www.directtv.com*) quando adicionou o canal National Geographic à sua programação de 2001. Como forma de promover seu serviço, a DIRECTV ofereceu três níveis de desconto para novos clientes. A empresa sentiu que isso incentivaria os novos clientes sensíveis a desconto a escolher os serviços da DIRECTV e teve sucesso na promoção de seus serviços.[1] ■

No exemplo da loja de departamentos, havia dois grupos de entrevistados (familiarizados e não-familiarizados), enquanto o exemplo da propensão para o desconto examinou três grupos (não-usuários, usuários médios e usuários freqüentes de descontos). Em ambos os estudos, as diferenças significativas inter-grupos foram encontradas usando variáveis previsoras múltiplas. Uma análise das diferenças entre grupos consiste no âmago do conceito básico da análise discriminante.

CONCEITO BÁSICO

A ***análise discriminante*** é uma técnica de análise de dados quando a variável dependente é categórica e as variáveis previsoras ou independentes têm natureza intervalar.[2] Por exemplo, a variável dependente pode ser a escolha de uma marca de computador pessoal (marca A, B ou C) e as variáveis independentes podem ser as avaliações dos atributos dos PCs em uma escala Likert de sete pontos. Os objetivos da análise discriminante são:

1. Estabelecer ***funções discriminantes***, ou combinações lineares das variáveis independentes ou previsoras, que melhor discriminem entre as categorias da variável dependente (grupos).
2. Verificar se existem diferenças significativas entre os grupos, em termos das variáveis previsoras.
3. Determinar as variáveis previsoras que mais contribuem para as diferenças entre grupos.
4. Enquadrar, ou classificar, os casos em um dos grupos com base nos valores das variáveis previsoras.
5. Avaliar a precisão da classificação.

análise discriminante
Técnica para analisar dados na pesquisa de marketing em que a variável dependente é categórica e as variáveis previsoras ou independentes têm natureza intervalar.

função discriminante
A combinação linear de variáveis independentes, estabelecida pela análise discriminante, que melhor discrimina entre as categorias da variável dependente.

análise discriminante de dois grupos
Técnica de análise discriminante em que a variável dependente tem duas categorias.

análise discriminante múltipla
Técnica de análise discriminante em que a variável dependente envolve três ou mais categorias.

As técnicas de análise discriminante são definidas pelo número de categorias que a variável dependente possui. Quando a variável dependente tem duas categorias, a técnica é conhecida como ***análise discriminante de dois grupos***. Quando estão em jogo três ou mais categorias, temos a ***análise discriminante múltipla***. A distinção principal é que, no caso de dois grupos, é possível deduzir apenas uma função discriminante, enquanto na análise discriminante múltipla podemos calcular mais de uma função.[3]

Os exemplos de análise discriminante são abundantes em pesquisa de marketing. Essa técnica pode ser aplicada para responder questões como:

- Em termos de características demográficas, qual é a diferença entre clientes freqüentes de um estabelecimento e clientes eventuais?
- Os usuários freqüentes, médios e eventuais de refrigerantes diferem quanto ao seu consumo de alimentos congelados?
- Que características psicográficas contribuem para estabelecer a diferença entre compradores de alimentos que se preocupam com o preço e compradores que não se preocupam com o preço?

[1] Anônimo, "DirecTv Adds National Geographic Channel," *Satellite News,* 24 (3) (January 15, 2001): 1; Donald R. Lichtenstein, Scot Burton e Richard G. Netemeyer, "An Examination of Deal Proneness Across Sales Promotion Types: A Consumer Segmentation Perspective," *Journal of Retailing,* 73 (2) (Summer 1997): 283-297; e Marvin A. Jolson, Joshua L. Wiener e Richard B. Rosecky, "Correlates of Rebate Proneness," *Journal of Advertising Research* (February/March 1987): 33-43.

[2] Pode-se encontrar uma discussão detalhada a respeito da análise discriminante em E. K. Kemsley, *Discriminant Analysis and Class Modeling of Spectroscopic Data* (New York: John Wiley & Sons, 1998); Jacques Tacq, *Multivariate Analysis Techniques in Social Science Research* (Thousand Oaks, CA: Sage Publications, 1997); e P. A. Lachenbruch, *Discriminant Analysis* (New York: Hafner Press, 1975). Para uma aplicação recente, ver Vipin Gupta, Paul J. Hanges e Peter Dorfman, "Cultural Clusters: Methodology and Findings," *Journal of World Business,* 37 (1) (Spring 2002): 11.

[3] Ver Richard Arnold Johnson e Dean W. Wichern, *Applied Multivariate Statistical Analysis* (Paramus, NJ: Prentice Hall, 2001); e W. R. Klecka, *Discriminant Analysis* (Beverly Hills, CA: Sage,1980).

- Os diversos segmentos do mercado diferem quanto a seus hábitos médios de consumo?
- Em termos de estilo de vida, quais as diferenças entre os clientes freqüentes de lojas de departamentos regionais e os de redes nacionais?
- Quais são as características dos consumidores que reagem positivamente às solicitações por mala direta?

RELAÇÃO ENTRE REGRESSÃO E ANOVA

A Tabela 18.1 exibe a relação entre análise discriminante, análise da variância (ANOVA) e análise de regressão. Ilustramos essa relação com um exemplo em que o pesquisador procura explicar o montante de seguro de vida contratado em função da idade e da renda. Os três procedimentos envolvem uma única variável dependente e variáveis independentes ou previsoras múltiplas. Entretanto, a natureza dessas variáveis difere. Em análise da variância e análise de regressão, a variável dependente é métrica ou intervalar (montante de seguro de vida contratado em dólares), enquanto na análise discriminante é categórica (montante de seguro de vida contratado classificado como alto, médio ou baixo). As variáveis independentes são categóricas no caso da análise da variância (idade e renda são classificadas como alta, média ou baixa), mas são métricas no caso de análise de regressão e análise discriminante (idade em anos e renda em dólares, ou seja, ambas medidas em uma escala de razão).

A análise discriminante de dois grupos, em que a variável dependente tem apenas duas categorias, está estreitamente relacionada com a análise de regressão múltipla. Nesse caso, a regressão múltipla, em que a variável dependente é codificada como uma variável *dummy* 0 ou 1, resulta em coeficientes de regressão parcial proporcionais aos coeficientes da função discriminante (ver a seção seguinte sobre o modelo de análise discriminante).

MODELO DE ANÁLISE DISCRIMINANTE

O ***modelo de análise discriminante*** envolve combinações lineares da seguinte forma:

$$D = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

onde

D = escore discriminante
bs = coeficiente ou peso discriminante
Xs = variável previsora ou independente

Os coeficientes, ou pesos (b), são estimados de modo que os grupos difiram tanto quanto possível em relação aos valores da função discriminante. Isso ocorre quando a razão da soma de quadrados entre grupos para a soma de quadrados dentro de grupos para os escores discriminantes for máxima. Qualquer outra combinação linear dos previsores resultará em uma razão menor. Os detalhes técnicos da estimativa constam no Apêndice 18.1. Várias estatísticas estão associadas à análise discriminante.

modelo de análise discriminante
Modelo estatístico no qual se baseia a análise discriminante.

ESTATÍSTICAS ASSOCIADAS À ANÁLISE DISCRIMINANTE

As estatísticas importantes associadas à análise discriminante incluem as seguintes:

Correlação canônica. A correlação canônica mede o alcance da associação entre os escores discriminantes e os grupos. É uma medida de associação entre a função discriminante isolada e o conjunto de variáveis *dummy* que definem a integração ao grupo.

Centróide. O centróide é a média dos valores dos escores discriminantes de um determinado grupo. Há tantos centróides quanto grupos – um para cada grupo. As médias para um grupo sobre todas as funções são os *centróides de grupos*.

Matriz de classificação. Por vezes também chamada de *matriz de confusão* ou *matriz de predição*, a matriz de classificação contém o total de casos classificados corretamente e mal classificados. Os casos classificados corretamente aparecem na diagonal, porque os grupos previstos e efetivos são os mesmos. Os elementos fora da diagonal representam casos que foram classificados incorretamente. A soma dos elementos da diagonal dividida pelo número total de casos representa a *proporção de acertos*.

TABELA 18.1
Semelhanças e diferenças entre ANOVA, regressão e análise discriminante

	ANOVA	Regressão	Análise discriminante
Semelhanças			
Número de variáveis dependentes	Uma	Uma	Uma
Número de variáveis independentes	Múltipla	Múltipla	Múltipla
Diferenças			
Natureza da variável dependente	Métrica	Métrica	Categórica
Natureza das variáveis independentes	Categórica	Métrica	Métrica

Coeficientes da função discriminante. Os coeficientes da função discriminante (não-padronizados) são os multiplicadores de variáveis quando as variáveis estão nas unidades de medida originais.

Escores discriminantes. Os coeficientes não-padronizados são multiplicados pelos valores das variáveis. Esses produtos são somados e adicionados ao termo constante, dando os escores discriminantes.

Autovalores. Para cada função discriminante, o autovalor é a razão da soma de quadrados entre grupos para a mesma soma dentro de grupos. Grandes autovalores implicam funções superiores.

Valores* F *e sua significância. São calculados pela ANOVA de um critério, com a variável agrupadora servindo como variável independente categórica. Cada previsor, por sua vez, serve como variável dependente métrica na ANOVA.

Médias de grupo e desvios-padrão de grupo. São calculados para cada previsor e para cada grupo.

Matriz de correlação combinada dentro de grupos. A matriz de correlação combinada dentro de grupos é calculada tomando-se a média das matrizes de covariância separadas para todos os grupos.

Coeficientes da função discriminante padronizada. Coeficientes da função discriminante usados como multiplicadores quando as variáveis foram padronizadas para média 0 e variância 1.

Correlações estruturais. Também chamadas de *cargas discriminantes*, as correlações estruturais representam as correlações simples entre os previsores e a função discriminante.

Matriz de correlação total. Se os casos forem tratados como se proviessem de uma única amostra, calculando-se então as correlações, obtém-se uma matriz de correlação total.

λ *de Wilks.* Por vezes chamado também de estatística U, o λ de Wilks para cada previsor é a razão da soma de quadrados dentro dos grupos para a soma total de quadrados. Seu valor varia entre 0 e 1. Grandes valores de λ (próximos de 1) indicam que as médias dos grupos não parecem ser diferentes umas das outras. Pequenos valores de λ (próximos de 0), entretanto, indicam que as médias de grupos parecem diferentes.

As suposições na análise discriminante são que cada grupo seja uma amostra de uma população normal multivariada e que todas as populações tenham a mesma matriz de covariância. Podemos entender melhor o papel dessas suposições e as estatísticas que acabamos de descrever estudando o procedimento para fazer análise discriminante.

COMO FAZER ANÁLISE DISCRIMINANTE

Os estágios envolvidos na realização de uma análise discriminante são formulação, estimação, determinação da significância, interpretação e validação (ver Figura 18.1). Esses estágios são discutidos e ilustrados dentro do contexto da análise discriminante de dois grupos. Mais adiante neste capítulo abordaremos a análise discriminante com mais de dois grupos.

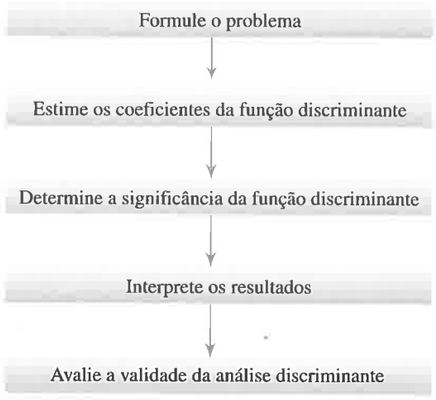

Figura 18.1 Como fazer análise discriminante.

Formular o problema

O primeiro passo na análise discriminante é formular o problema identificando os objetivos, a variável dependente e as variáveis independentes. A variável dependente deve consistir em duas ou mais categorias mutuamente excludentes e coletivamente exaustivas. Quando a variável dependente é escalonada em forma de intervalo ou razão, deve ser primeiramente convertida em categorias. Por exemplo, a atitude em relação a uma marca, medida em uma escala de sete pontos, poderia ser categorizada como desfavorável (1, 2, 3), neutra (4) ou favorável (5, 6, 7). Alternativamente, pode-se fazer o gráfico da distribuição da variável dependente e formar grupos de igual tamanho determinando os pontos de corte apropriados para cada categoria. As variáveis previsoras devem ser selecionadas com base em um modelo teórico ou em uma pesquisa prévia; no caso de pesquisa exploratória, a experiência do pesquisador deve orientar sua escolha.

O passo seguinte consiste em dividir a amostra em duas partes. Uma, chamada de amostra de estimação ou ***amostra de análise***, é usada para estimar a função discriminante. A outra parte, chamada de ***amostra de validação*** ou amostra retida (*holdout*), é reservada para validar a função discriminante. Quando a amostra é suficientemente grande, pode ser dividida ao meio. Uma das metades serve como amostra de análise e a outra é usada para validação. Invertem-se então os papéis das duas metades e repete-se a análise. Esse procedimento é chamado de dupla validação cruzada e é semelhante ao procedimento discutido na análise de regressão (Capítulo 17).

amostra de análise
Parte da amostra total que é utilizada para estimação da função discriminante.

amostra de validação
Parte da amostra total usada para verificar os resultados da amostra de estimação.

Com freqüência, a distribuição do número de casos nas amostras de análise e de validação segue a distribuição na amostra total. Por exemplo, se a amostra total contém 50% de clientes freqüentes e 50% de clientes eventuais, então as amostras de análise e de validação conterão cada uma 50% de clientes freqüentes e 50% de eventuais. Por outro lado, se amostra contivesse 25% de clientes freqüentes e 75% de clientes eventuais, as amostras de análise e de validação seriam selecionadas de modo a refletirem a mesma distribuição (25% *versus* 75%).

Finalmente, sugeriu-se que a validação da função discriminante se faça repetidamente. A cada vez, a amostra deve ser dividida em diferentes partes para análise e para validação. Deve-se estimar a função discriminante e fazer a análise de validação. Assim, a validação se baseia em várias tentativas. Foram sugeridos também métodos mais rigorosos.[4]

Para melhor ilustrar a análise discriminante de dois grupos, consideremos um exemplo. Suponha que queiramos determinar as características relevantes de famílias que freqüentaram uma estação de férias durante os últimos dois anos. Obtiveram-se dados de uma amostra pré-teste de 42 residências. Dessas, 30 residências mostradas na Tabela 18.2 foram incluídas na amostra de análise e as 12 restantes, mostradas na Tabela 18.3, integraram a amostra de validação. Os moradores que visitaram a estação durante os dois últimos anos foram codificados como 1; os que não visitaram foram codificados como 2 (VISITA). Ambas as amostras de análise e de validação foram equilibradas em termos de VISITA. Como se pode ver, a amostra de análise contém 15 residências em cada categoria, enquanto a amostra de validação contém 6 em cada categoria. Obtiveram-se também dados sobre a renda familiar anual (RENDA), atitude em relação à viagem (VIAGEM, medida em uma escala de nove pontos), importância atribuída às férias da família (FÉRIAS, medida em uma escala de nove pontos), tamanho da residência (TAMANHO) e idade do chefe da casa (IDADE).

Estimar os coeficientes da função discriminante

Uma vez identificada a amostra de análise, como na Tabela 18.2, podemos estimar os coeficientes da função discriminante. Dispomos de dois amplos métodos de abordagem. O ***método direto*** consiste em estimar a função discriminante de modo que todos os previsores sejam incluídos simultaneamente. Nesse caso, todas as variáveis independentes são incluídas, independentemente de seu poder discriminatório. Esse método é adequado quando, com base em uma pesquisa prévia ou em um modelo teórico, o pesquisador deseja que a discriminação esteja baseada em todos os previsores. Uma abordagem alternativa é o método passo a passo. Na ***análise discriminante passo a passo***, as variáveis previsoras são introduzidas seqüencialmente, com base em sua capacidade de discriminar entre os grupos. Esse método, detalhado mais adiante, é adequado quando o pesquisador deseja selecionar um subconjunto dos previsores para inclusão na função discriminante.

A Tabela 18.4 apresenta os resultados de uma análise discriminante de dois grupos sobre os dados da Tabela 18.2, utilizando um conhecido programa de computador. Pode-se obter uma percepção intuitiva dos resultados examinando as médias e os desvios-padrão dos grupos. Parece que os dois grupos são mais separados em termos de renda do que em termos de outras variáveis. Parece haver mais separação na importância atribuída às férias da família do que à atitude em relação à viagem. A diferença entre os dois grupos, no tocante à idade do chefe da família, é pequena, e o desvio-padrão dessa variável é grande.

[4] Philip Hans Franses, "A Test for the Hit Rate in Binary Response Models," *International Journal of Market Research,* 42 (2) (Spring 2000): 239-245; Vincent-Watne Mitchell, "How to Identify Psychographic Segments: Part 2," *Marketing Intelligence & Planning,* 12 (7) (1994): 11-16; e M. R. Crask e W. D. Perreault, Jr., "Validation of Discriminant Analysis in Marketing Research," *Journal of Marketing Research,* 14 (February 1977): 60-68.

A análise discriminante de dois grupos pode ser usada para determinar as principais características de famílias que visitaram uma estação de férias.

TABELA 18.2

Informações sobre visitas a estações de férias: amostra de análise

Número	Visitas à estação	Renda familiar anual (US$ 000)	Atitude em relação à viagem da família	Importância atribuída às férias da família	Tamanho da residência	Idade do chefe da família	Valor gasto com as férias
1	1	50,2	5	8	3	43	M (2)
2	1	70,3	6	7	4	61	A (3)
3	1	62,9	7	5	6	52	A (3)
4	1	48,5	7	5	5	36	B (1)
5	1	52,7	6	6	4	55	A (3)
6	1	75,0	8	7	5	68	A (3)
7	1	46,2	5	3	3	62	M (2)
8	1	57,0	2	4	6	51	M (2)
9	1	64,1	7	5	4	57	A (3)
10	1	68,1	7	6	5	45	A (3)
11	1	73,4	6	7	5	44	A (3)
12	1	71,9	5	8	4	64	A (3)
13	1	56,2	1	8	6	54	M (2)
14	1	49,3	4	2	3	56	A (3)
15	1	62,0	5	6	2	58	A (3)
16	2	32,1	5	4	3	58	B (1)
17	2	36,2	4	3	2	55	B (1)
18	2	43,2	2	5	2	57	M (2)
19	2	50,4	5	2	4	37	M (2)
20	2	44,1	6	6	3	42	M (2)
21	2	38,3	6	6	2	45	B (1)
22	2	55,0	1	2	2	57	M (2)
23	2	46,1	3	5	3	51	B (1)
24	2	35,0	6	4	5	64	B (1)
25	2	37,3	2	7	4	54	B (1)
26	2	41,8	5	1	3	56	M (2)
27	2	57,0	8	3	2	36	M (2)
28	2	33,4	6	8	2	50	B (1)
29	2	37,5	3	2	3	48	B (1)
30	2	41,3	3	3	2	42	B (1)

TABELA 18.3

Informações sobre visitas a uma estação de férias: amostra de validação

Número	Visitas à estação	Renda familiar anual (US$ 000)	Atitude em relação à viagem da família	Importância atribuída às férias da família	Tamanho da residência	Idade do chefe da família	Valor gasto com as férias
1	1	50,8	4	7	3	45	M (2)
2	1	63,6	7	4	7	55	A (3)
3	1	54,0	6	7	4	58	M (2)
4	1	45,0	5	4	3	60	M (2)
5	1	68,0	6	6	6	46	A (3)
6	1	62,1	5	6	3	56	A (3)
7	2	35,0	4	3	4	54	B (1)
8	2	49,6	5	3	5	39	B (1)
9	2	39,4	6	5	3	44	A (3)

(*continua*)

TABELA 18.3

Informações sobre visitas a uma estação de férias: amostra de validação

Número	Visitas à estação	Renda familiar anual (US$ 000)	Atitude em relação à viagem da família	Importância atribuída às férias da família	Tamanho da residência	Idade do chefe da família	Valor gasto com as férias
10	2	37,0	2	6	5	51	B (1)
11	2	54,5	7	3	3	37	M (2)
12	2	38,2	2	2	3	49	B (1)

TABELA 18.4

Resultados da análise discriminante de dois grupos

Médias dos grupos

Visita	Renda	Viagem	Férias	Tamanho da residência	Idade
1	60,52000	5,40000	5,80000	4,33333	53,73333
2	41,91333	4,33333	4,06667	2,80000	50,13333
Total	51,21667	4,86667	4,93333	3,56667	51,93333

Desvios-padrão dos grupos

1	9,83065	1,91982	1,82052	1,23443	8,77062
2	7,55115	1,95180	2,05171	0,94112	8,27101
Total	12,79523	1,97804	2,09981	1,33089	8,57395

Matriz de correlação combinada dentro de grupos

Visita	Renda	Viagem	Férias	Tamanho da residência	Idade
RENDA	1,00000				
VIAGEM	0,19745	1,00000			
FÉRIAS	0,09148	0,08434	1,00000		
TAMANHO DA RESIDÊNCIA	0,08887	– 0,01681	0,07046	1,00000	
IDADE	– 0,01431	– 0,19709	0,01742	– 0,04301	1,00000

λ de Wilks (estatística *U*) e razão *F* univariada com 1 e 28 graus de liberdade

Variável	λ de Wilks	F	Significância
RENDA	0,45310	33,80	0,0000
VIAGEM	0,92479	2,277	0,1425
FÉRIAS	0,82377	5,990	0,0209
TAMANHO DA RESIDÊNCIA	0,65672	14,64	0,0007
IDADE	0,95441	1,338	0,2572

Funções discriminantes canônicas

Função	Autovalor	Percentagem de variação	Percentagem acumulada	Correlação canônica	Função após	λ de Wilks	Qui-quadrado	gl	Sig.
					: 0	0,3589	26,130	5	0,0001
1*	1,7862	100,00	100,00	0,8007	:				

*Assinala as funções discriminantes canônicas restantes na análise.

Coeficientes padronizados da função discriminante canônica

	Func 1
RENDA	0,74301
VIAGEM	0,09611

(*continua*)

TABELA 18.4

Resultados da análise discriminante de dois grupos

FÉRIAS	0,23329
TAMANHO DA RESIDÊNCIA	0,46911
IDADE	0,20922

MATRIZ ESTRUTURAL

Correlações combinadas dentro de grupos entre variáveis discriminantes e funções discriminantes canônicas (variáveis ordenadas pelo tamanho da correlação dentro da função)

	FUNC 1
RENDA	0,82202
TAMANHO DA RESIDÊNCIA	0,54096
FÉRIAS	0,34607
VIAGEM	0,21337
IDADE	0,16354

COEFICIENTES NÃO-PADRONIZADOS DA FUNÇÃO DISCRIMINANTE CANÔNICA

	FUNC 1
RENDA	0,8476710E-01
VIAGEM	0,4964455E-01
FÉRIAS	0,1202813
TAMANHO DA RESIDÊNCIA	0,4273893
IDADE	0,2454380E-01
(constante)	–7,975476

FUNÇÕES DISCRIMINANTES CANÔNICAS CALCULADAS PARA MÉDIAS DE GRUPOS (CENTRÓIDES DE GRUPOS)

GRUPO	FUNC1
1	1,29118
2	– 1,29118

RESULTADOS DE CLASSIFICAÇÃO

			AFILIAÇÃO PREVISTA AO GRUPO		TOTAL
		VISITA	1	2	
Original	Contagem	1	12	3	15
		2	0	15	15
	%	1	80,0	20,0	100,0
		2	0,0	100,0	100,0
Validação cruzada	Contagem	1	11	4	15
		2	2	13	15
	%	1	73,5	26,7	100,0
		2	13,3	86,7	100,0

[a] A validação cruzada é feita somente para os casos que estão na análise. Na validação cruzada, cada caso é classificado pela função deduzida de todos os outros casos.

[b] 90,0% dos casos agrupados originais classificados corretamente.

[c] 80,0% dos casos agrupados por validação cruzada classificados corretamente.

RESULTADOS DE CLASSIFICAÇÃO PARA CASOS NÃO-SELECIONADOS PARA USO NA ANÁLISE

	GRUPO EFETIVO	NÚMERO DE CASOS	AFILIAÇÃO PREVISTA AO GRUPO 1	2
Grupo	1	6	4	2
			66,7%	33,3%
Grupo	2	6	0	6
			0,0%	100,0%

Percentagem de casos agrupados classificados corretamente: 83,33%

método direto

Método de abordagem da análise discriminante que envolve a estimação da função discriminante de forma que todos os previsores sejam incluídos simultaneamente.

análise discriminante passo a passo

Procedimento de análise discriminante em que os previsores são introduzidos seqüencialmente com base em sua capacidade de discriminar entre os grupos.

A matriz de correlação combinada dentro de grupos indica baixas correlações entre os previsores. É improvável que a multicolinearidade venha a constituir um problema. A significância da razão F univariada indica que, quando os previsores são analisados individualmente, apenas a renda, a importância das férias e o tamanho da residência distinguem significativamente os que freqüentam uma estação de férias dos que não freqüentam.

Como há dois grupos, estima-se apenas uma função discriminante. O autovalor associado a essa função é 1,7862 e responde por 100% da variância explicada. A correlação canônica associada a essa função é 0,8007. O quadrado dessa correlação, $(0,8007)^2 = 0,64$, indica que 64% da variância na variável dependente (VISITA) são explicados por esse modelo.

Determinar a significância da função discriminante

Não teria sentido interpretarmos a análise se as funções discriminantes estimadas não fossem significativas. Pode-se testar estatisticamente a hipótese nula de que as médias de todas as funções discriminantes em todos os grupos sejam iguais. No SPSS, esse teste se baseia no λ de Wilks. Testando-se simultaneamente várias funções (como no caso da análise discriminante múltipla), a estatística λ de Wilks é o produto dos λ univariados de cada função. Estima-se o nível de significância com base em uma transformação qui-quadrado da estatística. Ao testar a significância no exemplo da estação de férias (ver Tabela 18.4), vê-se que o λ de Wilks associado à função é 0,3589, que se transforma em um qui-quadrado 26,13 com 5 graus de liberdade; é significativo acima do nível de 0,05. No SAS, calcula-se uma estatística F com base em uma aproximação da distribuição da razão de verossimilhança. Não dispomos de teste em MINITAB. Se a hipótese nula for rejeitada, o que indica discriminação significativa, pode-se passar à interpretação dos resultados.[5]

Interpretar os resultados

A interpretação dos pesos, ou coeficientes, da discriminante é similar à do caso da análise de regressão múltipla. O valor do coeficiente para um determinado previsor depende dos outros previsores incluídos na função discriminante. Os sinais dos coeficientes são arbitrários, mas indicam que valores da variável resultam em grandes ou pequenos valores da função, associando-os a grupos particulares.

Dada a multicolinearidade nas variáveis previsoras, não há medida não-ambígua da importância relativa dos previsores na discriminação entre os grupos.[6] Com essa precaução em mente, podemos ter alguma idéia da importância relativa das variáveis, examinando a magnitude absoluta dos coeficientes padronizados da função discriminante. De modo geral, os previsores com coeficientes padronizados relativamente grandes contribuem mais para o poder discriminatório da função, em comparação com previsores com coeficientes menores.

Podemos também ter idéia da importância relativa dos previsores examinando as correlações estruturais, também chamadas de *cargas canônicas* ou *cargas discriminantes*. Essas correlações simples entre cada previsor e a função discriminante representam a variância que o previsor compartilha com a função. Tal como os coeficientes padronizados, essas correlações também devem ser interpretadas com cautela.

É conveniente e instrutivo examinar os coeficientes padronizados da função discriminante para o caso do exemplo da estação de férias. Dadas as baixas intercorrelações entre os previsores, podem-se utilizar com cautela as magnitudes dos coeficientes padronizados para sugerir que a renda é o previsor mais importante na discriminação entre os grupos, seguida pelo tamanho da residência e pela importância atribuída às férias da família. Obtém-se a mesma informação do exame das correlações estruturais. Relacionam-se a seguir, em ordem de grandeza, essas correlações simples entre os previsores e a função discriminante.

Os coeficientes não-padronizados da função discriminante também são dados. Podem ser aplicados aos valores brutos das variáveis no conjunto retido para fins de classificação. São apresentados também os centróides dos grupos, que dão o valor da função discriminante calculado nas médias dos grupos. O grupo 1 (os que visitaram uma estação de férias) tem valor positivo, enquanto o grupo 2 tem valor negativo igual. Os sinais dos coeficientes associados a todos os previsores são positivos. Isso sugere que a freqüência da família a uma estação pode resultar de uma renda familiar mais alta, do tamanho da residência, da importância atribuída às férias, da atitude em relação à viagem e da idade. Seria interessante traçar um perfil dos dois grupos em termos dos previsores mais importantes: renda, tamanho da residência e importância das férias. A Tabela 18.4 dá os valores dessas três variáveis para os dois grupos.

O exemplo a seguir apresenta mais uma ilustração da determinação da importância relativa dos previsores.

PESQUISA REAL

Os vendedores satisfeitos permanecem

Uma pesquisa realizada em 2001 interrogou executivos a respeito da preocupação de contratar e manter funcionários durante o

[5] Falando em termos estritos, antes de testar a igualdade das médias de grupo, é preciso testar as matrizes de covariância de grupo. O teste chamado de "Box's M" pode ser usado para esse propósito. Se a igualdade das matrizes de covariância de grupo for rejeitada, os resultados da análise discriminante devem ser interpretados com cuidado. Nesse caso, diminui a potência do teste para a igualdade das médias de grupo.

[6] Ver Nessim Hanna, "Brain Dominance and the Interpretation of Advertising Messages," *International Journal of Commerce & Management, 9* (3/4) (1999): 19-32; Lillian Fok, John P. Angelidis, Nabil A. Ibrahim e Wing M. Fok, "The Utilization and Interpretation of Multivariate Statistical Techniques in Strategic Management," *International Journal of Management,* 12 (4) (December 1995): 468-481; e D. G. Morrison, "On the Interpretation of Discriminant Analysis," *Journal of Marketing Research, 6* (May 1969): 156-163.

RESULTADO DA ANÁLISE DISCRIMINANTE

Variável	*Coeficientes*	*Coeficientes padronizados*	*Cargas canônicas*
1. Trabalho[a]	0,0903	0,3910	0,5446
2. Promoção[a]	0,0288	0,1515	0,5044
3. Segurança no emprego	0,1567	0,1384	0,4958
4. Relações com os clientes[b]	0,0086	0,1751	0,4906
5. Avaliação da companhia	0,4059	0,3240	0,4824
6. Trabalho em equipe[b]	0,0018	0,0365	0,4651
7. Desempenho global[b]	– 0,0148	– 0,3252	0,4518
8. Gerenciamento tempo-território[b]	0,0126	0,2899	0,4496
9. Vendas produzidas[b]	0,0059	0,1404	0,4484
10. Habilidade na apresentação[b]	0,0118	0,2526	0,4387
11. Informação técnica[b]	0,0003	0,0065	0,4173
12. Adicionais sobre salário[a]	0,0600	0,1843	0,3788
13. Quota atingida[b]	0,0035	0,2915	0,3780
14. Gerência[a]	0,0014	0,0138	0,3571
15. Coleta de informações[b]	– 0,0146	– 0,3327	0,3326
16. Família[c]	– 0,0684	– 0,3408	– 0,3221
17. Gerente de vendas[a]	– 0,0121	– 0,1102	0,2909
18. Colega de trabalho[a]	0,0225	0,0893	0,2671
19. Cliente[c]	– 0,0625	– 0,2797	– 0,2602
20. Família[d]	0,0473	0,1970	0,2180
21. Emprego[d]	0,1378	0,5312	0,2119
22. Emprego[c]	0,0410	0,5475	– 0,1029
23. Cliente[d]	– 0,0060	– 0,0255	0,1004
24. Gerente de vendas[c]	– 0,0365	– 0,2406	– 0,0499
25. Gerente de vendas[d]	– 0,0606	– 0,3333	0,0467
26. Cliente[a]	– 0,0338	– 0,1488	0,0192

Nota: A ordem de importância da classificação se baseia na magnitude das cargas canônicas.
[a] Satisfação
[b] Desempenho
[c] Ambigüidade
[d] Conflito

conturbado clima econômico vigente naquele período. Relatou-se que 85% dos entrevistados se preocupava com a contratação de funcionários e 81% afirmaram que estavam preocupados em manter seus funcionários. Quando a economia está em baixa, a rotatividade é rápida. Em termos gerais, se uma organização deseja reter seus funcionários, ela precisa aprender por que as pessoas abandonam seus empregos e por que outros ficam e estão satisfeitos com seu trabalho. Utilizou-se a análise discriminante para determinar os fatores que explicam a diferença entre vendedores que deixam um grande fabricante de computadores e os que permanecem. As variáveis independentes foram a avaliação da companhia, segurança no emprego, sete itens de satisfação no emprego, quatro aspectos do conflito de atribuições, quatro aspectos da ambigüidade de atribuições e nove medidas de desempenho em vendas. A variável dependente foi a dicotomia entre os que permaneceram na empresa e os que saíram. A correlação canônica, um índice de discriminação ($R = 0{,}4572$), foi significante (λ de Wilks = 0,7909, $F(26,173) = 1{,}7588$, $p = 0{,}0180$). Esse resultado mostrou que as variáveis discriminavam entre os que saíram da empresa e os que nela permaneceram.

Os resultados da introdução simultânea de todas as variáveis na análise discriminante foram apresentados na tabela a seguir. A importância da ordem de classificação, tal como determinada pela magnitude relativa das correlações da estrutura, aparece na primeira coluna. A satisfação com o emprego e as oportunidades de promoção foram dois dos discriminantes mais importantes, seguidos pela segurança no emprego. Os que permaneceram na empresa achavam o emprego mais motivador, mais satisfatório, mais desafiador e mais interessante do que aqueles que saíram.[7]

Observe que, nesse exemplo, a promoção foi identificada como a segunda variável mais importante, com base nas correlações estruturais. Todavia, não é a segunda variável mais importante com base na magnitude absoluta dos coeficientes padronizados da função discriminante. Essa anomalia resulta da multicolinearidade. ■

[7] Robert J. Sahl, "Retention Reigns as Economy Suffers Drought," *Workspan,* 44 (11) (November 2001): 6-8; Jon M. Hawes, C. P. Rao e Thomas L. Baker, "Retail Salesperson Attributes and the Role of Dependability in the Selection of Durable Goods," *Journal of Personal Selling & Sales Management,* 13 (4) (Fall 1993): 61-71; Edward F. Fern, Ramon A. Avila e Dhruv Grewal, "Salesforce Turnover: Those Who Left and Those Who Stayed," *Industrial Marketing Management* (1989): 1-9.

Outra ajuda na interpretação dos resultados da análise discriminante consiste em estabelecer um ***perfil das características*** para cada grupo, descrevendo cada um deles em termos das médias grupais para as variáveis previsoras. Se os previsores importantes foram identificados, então uma comparação das médias grupais dessas variáveis pode ajudar a compreender as diferenças intergrupos. Entretanto, antes de poder interpretar com segurança quaisquer constatações, é necessário validar os resultados.

perfil das características
Um auxílio na interpretação dos resultados da análise discriminante por meio da descrição de cada grupo em termos das médias de grupo para as variáveis previsoras.

Avaliar a validade da análise discriminante

Muitos programas de computador, como o SPSS, oferecem uma opção de validação cruzada que deixa de fora um entrevistado. Nessa opção, o modelo discriminante é estimado novamente o número de vezes igual ao de entrevistados na amostra. Cada modelo estimado novamente deixa de fora um entrevistado e o modelo é usado para fazer a previsão para esse entrevistado. Quando uma grande amostra de validação não é possível, isso oferece uma noção da robusteza da estimação usando cada respondente, na sua vez, como retentor.

Conforme explicamos anteriormente, sempre que possível, os dados devem ser divididos aleatoriamente em duas subamostras. Uma delas, a amostra de análise, é usada para estimar a função discriminante; a amostra de validação serve para desenvolver a matriz de classificação. Os pesos discriminantes, estimados com auxílio da amostra de análise, são multiplicados pelos valores das variáveis previsoras na amostra retida para gerar escores discriminantes para os casos nessa amostra. Os casos são então atribuídos a grupos com base em seus escores discriminantes e em uma regra de decisão adequada. Por exemplo, na análise discriminante de dois grupos, um caso será atribuído ao grupo cujo centróide é o mais próximo. Pode-se então determinar a ***proporção de acertos***, ou percentagem de casos classificados corretamente, somando os elementos da diagonal e dividindo o resultado pelo número total de casos.[8]

proporção de acertos
Percentagem de casos classificados corretamente pela análise discriminante.

É conveniente comparar a percentagem dos casos classificados corretamente pela análise discriminante com a que seria obtida aleatoriamente. Quando os grupos têm tamanhos iguais, a percentagem de classificação aleatória é 1 dividido pelo número de grupos. Qual o grau de melhora a ser esperado, então, em comparação com o acaso? Não dispomos de diretrizes gerais, embora alguns autores tenham sugerido que a precisão de classificação obtida pela análise discriminante deva ser ao menos 25% maior do que a obtida aleatoriamente.[9]

A maioria dos programas de análise discriminante também estima uma matriz de classificação com base na amostra de análise. Como eles capitalizam na variação aleatória dos dados, tais resultados são invariavelmente melhores do que a classificação obtida na amostra retida.[10]

A Tabela 18.4, do exemplo da estação de férias, também apresenta os resultados da classificação baseada na amostra de análise. A proporção de acertos, ou percentagem de casos classificados corretamente, é (12 + 15)/30 = 0,90, ou 90%. Pode-se suspeitar que essa proporção esteja artificialmente inflada, pois os dados usados para estimação foram também usados para validação. A realização da análise de classificação sobre um conjunto retido de dados independentes tem como resultado uma matriz de classificação com proporção de acertos ligeiramente inferior a (4+6)/12 = 0,833, ou 83,3% (ver Tabela 18.4). Dados dois grupos de tamanhos iguais, pelo acaso esperaríamos uma razão de acerto de 1/2 = 0,50, ou 50%. Logo, a melhora em relação à chance é superior a 25% e a validade da análise discriminante é considerada satisfatória.

PESQUISA REAL

Caseiros e telemaníacos

Foi utilizada a análise discriminante de dois grupos para avaliar a intensidade de cada uma das cinco dimensões usadas na classificação dos indivíduos como usuários ou não-usuários de TV. O procedimento foi apropriado para o objetivo pretendido em razão da natureza dos grupos categóricos predefinidos (usuários e não-usuários) e das escalas intervalares usadas para gerar escores fatoriais individuais.

Foram criados dois grupos iguais de 185 consumidores idosos (total $n = 370$). Estimou-se a equação discriminante para análise utilizando-se uma subamostra de 142 entrevistados da amostra de 370. Dos entrevistados restantes, 198 foram utilizados como subamostra de validação em uma validação cruzada da equação. Foram excluídos da análise 30 entrevistados em virtude de valores discriminantes faltantes.

A correlação canônica para a função discriminante foi 0,4291, significativa ao nível $p < 0{,}0001$. O autovalor foi 0,2257. A tabela resume os coeficientes discriminantes canônicos padronizados. A função discriminante explica uma parte

[8] Para a validação da análise discriminante, ver Werner J. Reinartz e V. Kumar, "On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing," *Journal of Marketing,* 64 (4) (October 2000): 17-35.

[9] Joseph F. Hair, Jr., Ralph E. Anderson, Ronald L. Tatham e William C. Black, *Multivariate Data Analysis with Readings,* 5th ed. (Englewood Cliffs, NJ: Prentice-Hall, Inc., 1998). Ver também J. J. Glen, "Classification Accuracy in Discriminant Analysis: A Mixed Integer Programming Approach," *The Journal of the Operational Research Society,* 52 (3) (March 2001): 328.

[10] Anônimo, "Interactive TV Growth Set to Erupt over Next Five Years," *Satellite News,* 24 (2) (January 8, 2001): 1; e Don R. Rahtz, M. Joseph Sirgy e Rustan Kosenko, "Using Demographics and Psychographic Dimensions to Discriminate Between Mature Heavy and Light Television Users: An Exploratory Analysis," in Kenneth D. Bahn, Ed., *Developments in Marketing Science,* Vol. 11 (Blacksburg, VA: Academy of Marketing Science, 1988): 2-7.

substancial da variância. Além disso, como a tabela mostra, a dimensão "orientação doméstica" deu uma contribuição bastante forte para a classificação dos indivíduos como usuários ou não-usuários da televisão. "Moral, segurança, saúde" e "respeito" também contribuíram significativamente. Não parece ter havido grande contribuição do fator "social".

O procedimento de validação cruzada utilizando a função discriminante da amostra de análise apoiou a afirmação de que as dimensões ajudaram os pesquisadores a discriminar entre usuários e não-usuários de televisão. Como a tabela mostra, a função discriminante foi bem sucedida ao classificar 75,76% dos casos. Isso sugere que a consideração das dimensões identificadas ajuda os pesquisadores de mercado a entenderem o mercado dos idosos. Embora seja muito importante que os pesquisadores conheçam e compreendam o mercado dos idosos, os membros da Geração X (aqueles que nasceram entre 1961 e 1981) também são um grupo que não deve ser subestimado pelos profissionais. Devido a avanços tecnológicos com a TV e a Internet, uma forma revolucionária de TV interativa (iTV) foi criada. Até 2003, as previsões afirmavam que os serviços de iTV estariam inteiramente implementados e operacionais, combinando a Internet e a radiodifusão com programas de *software* e componentes de *hardware* para oferecer aos consumidores acesso à Internet, compras *on-line*, *download* de músicas e um programa de radiodifusão interativo em toda sua televisão. Até 2006, prevê-se que 61,5 milhões de pessoas estarão interagindo com seus televisores através da iTV. Com uma previsão de aparência tão próspera para a iTV, quem seria melhor para analisar essa forma revolucionária de televisão que a Geração X? A análise discriminante pode, mais uma vez, ser usada para determinar quem são, da Geração X, os usuários e não-usuários da iTV e para vender serviços de iTV com êxito.[10] ■

A extensão da análise discriminante de dois grupos para a análise discriminante múltipla envolve etapas similares.

ANÁLISE DISCRIMINANTE MÚLTIPLA

Formular o problema

Os dados apresentados nas Tabelas 18.2 e 18.3 também podem ser utilizados para ilustrar a análise discriminante de três grupos. Na última coluna dessas tabelas, as residências são classificadas em três categorias, com base na despesa da família com férias (alta, média, baixa). Enquadram-se dez residências em cada categoria. A questão de interesse é se as casas em que se gastam grandes, médias ou pequenas quantias (QUANTIA) em suas férias podem ser diferenciadas em termos de renda familiar anual (RENDA), atitude em relação à viagem (VIAGEM), importância atribuída às férias (FÉRIAS), tamanho da residência (TAMANHO) e idade do chefe da família (IDADE).[11]

[10] Anônimo, "Interactive TV Growth Set to Erupt over Next Five Years," *Satellite News,* 24 (2) (January 8, 2001): 1; e Don R. Rahtz, M. Joseph Sirgy e Rustan Kosenko, "Using Demographics and Psychographic Dimensions to Discriminate Between Mature Heavy and Light Television Users: An Exploratory Analysis," in Kenneth D. Bahn, Ed., *Developments in Marketing Science,* Vol. 11 (Blacksburg, VA: Academy of Marketing Science, 1988): 2-7.

RESUMO DA ANÁLISE DISCRIMINANTE

Coeficientes padronizados da função discriminante canônica

Moral	0,27798
Segurança e saúde	0,39850
Orientação doméstica	0,77496
Respeito	0,32069
Social	– 0,01996

Resultados da classificação de casos selecionados para uso na análise

		Afiliação prevista ao grupo	
Grupo efetivo	***Número de casos***	***Não-usuários***	***Usuários***
Não-usuários de TV	77	56 72,7%	21 27,3%
Usuários de TV	65	24 36,9%	41 63,1%

Percentagem de casos agrupados corretamente classificados: 68,31%.

Resultados da classificação para casos usados para validação cruzada

		Afiliação prevista ao grupo	
Grupo efetivo	***Número de casos***	***Não-usuários***	***Usuários***
Não-usuários de TV	108	85 78,7%	23 21,3%
Usuários de TV	90	25 27,8%	65 72,2%

Percentagem de casos agrupados classificados corretamente: 75,76%

Estimar os coeficientes da função discriminante

A Tabela 18.5 apresenta os resultados da estimação de análise discriminante de três grupos. Um exame das médias dos grupos indica que a renda parece separar os grupos mais do que qualquer outra variável. Há alguma separação em viagem e férias. Os grupos 1 e 2 estão muito próximos em termos de tamanho da casa e idade. A idade acusa grande desvio-padrão em relação à separação entre os grupos. A matriz combinada de correlação dentro dos grupos indica alguma correlação de férias e tamanho da casa com a renda. A idade apresenta alguma correlação negativa com a viagem. Todavia, essas correlações estão do lado mais baixo, o que indica que, embora a multicolinearidade possa causar alguma preocupação, não constitui problema sério. A significância atribuída às razões F univariadas indica que, quando os previsores são analisados individualmente, apenas a renda e as viagens são significativas na diferenciação dos dois grupos.

Na análise discriminante múltipla, se há G grupos, podem-se estimar $G - 1$ funções discriminantes se o número de previsores for maior do que esse número. De modo geral, com G grupos e k previsores, é possível estimar até $G - 1$, ou k, funções discriminantes (o que for menor). A primeira função tem a maior razão entre a soma de quadrados entre grupos e a soma de quadrados dentro de grupos. A segunda função, não-correlacionada com a primeira, tem a segunda razão mais alta, e assim por diante. Não obstante, nem todas as funções podem ser estatisticamente significativas.

Como há três grupos, só podemos extrair, no máximo, duas funções. O autovalor associado à primeira função é 3,8190, e essa função responde por 93,93% da variância explicada. Como o autovalor é grande, a primeira função tende a ser superior. A segunda função tem um autovalor pequeno, 0,2469, e responde apenas por 6,07% da variância explicada.

Determinar a significância da função discriminante

Para testar a hipótese nula de iguais centróides de grupos, devemos considerar ambas as funções simultaneamente. É possível testar as médias das funções sucessivamente testando primeiro todas as médias simultaneamente. A seguir, exclui-se uma função de cada vez, testando-se as médias das funções restantes em cada passo. Na Tabela 18.5, o zero abaixo da "função após" indica que nenhuma função foi removida. O valor de λ de Wilks é 0,1644, que se transforma em um qui-quadrado de 44,831 com 10 graus de liberdade, significativo acima do nível de 0,05. Assim, as duas funções juntas discriminam significativamente entre os três grupos. Todavia, quando se remove a primeira função, o λ de Wilks associado à segunda função é 0,8020, que não é significativo ao nível de 0,05. Portanto, a segunda função não contribui significativamente para a diferença dos grupos.

Interpretar os resultados

A interpretação dos resultados é facilitada por um exame dos coeficientes padronizados da função discriminante, pelas correlações estruturais e por certos gráficos. Os coeficientes padronizados indicam um grande coeficiente para a renda na função 1, enquanto a função 2 tem coeficientes relativamente maiores para viagens, férias e idade. Chegamos a conclusão similar mediante exame da matriz estrutural (ver Tabela 18.5). Para ajudar a interpretar as funções, as variáveis com grandes coeficientes para uma função particular são agrupadas juntas. Esses equipamentos são assinalados com asteriscos. Assim, renda e tamanho da casa têm asteriscos para a função 1, porque essas variáveis têm coeficientes maiores para a função 1 do que para a função 2. Essas variáveis estão associadas essencialmente com a função 1. Por outro lado, viagens, férias e idade estão predominantemente associadas à função 2, tal como indicado pelos asteriscos.

[11] Richard A. Johnson e Dean W. Wichem, *Applied Multivariate Statistical Analysis,* 5th ed. (Upper Saddle River, NJ: Prentice-Hall, 2002). Para uma aplicação recente, ver Werner J. Reinartz e V. Kumar, "On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing," *Journal of Marketing,* 64 (4) (October 2000): 17-35.

TABELA 18.5

Resultados da análise discriminante de três grupos

MÉDIAS DE GRUPO QUANTIA	RENDA	VIAGENS	FÉRIAS	TAMANHO DA CASA	IDADE
1	38,57000	4,50000	4,70000	3,10000	50,30000
2	50,11000	4,00000	4,20000	3,40000	49,50000
3	64,97000	6,10000	5,90000	4,20000	56,00000
Total	51,21667	4,86667	4,93333	3,56667	51,93333
DESVIOS-PADRÃO DE GRUPOS					
1	5,29718	1,71594	1,88856	1,19722	8,09732
2	6,00231	2,35702	2,48551	1,50555	9,25263
3	8,61434	1,19722	1,66333	1,13529	7,60117
Total	12,79523	1,97804	2,09981	1,33089	8,57395

(*continua*)

TABELA 18.5

Resultados da análise discriminante de três grupos

MATRIZ DE CORRELAÇÃO COMBINADA DENTRO DE GRUPOS

	RENDA	VIAGENS	FÉRIAS	TAMANHO DA CASA	IDADE
RENDA	1,00000				
VIAGENS	0,05120	1,00000			
FÉRIAS	0,30681	0,03588	1,00000		
TAMANHO DA CASA	0,38050	0,00474	0,22080	1,00000	
IDADE	– 0,20939	– 0,34022	– 0,01326	– 0,02512	1,00000

λ de Wilks (estatística *U*) e razão *F* univariada com 2 e 27 graus de liberdade

VARIÁVEL	λ DE WILKS	*F*	SIGNIFICÂNCIA
RENDA	0,26215	38,00	0,0000
VIAGENS	0,78790	3,634	0,0400
FÉRIAS	0,88060	1,830	0,1797
TAMANHO DA CASA	0,87411	1,944	0,1626
IDADE	0,88214	1,804	0,1840

FUNÇÕES DISCRIMINANTES CANÔNICAS

FUNÇÃO	AUTOVALOR	PERCENTAGEM DE VARIAÇÃO	PERCENTAGEM ACUMULADA	CORRELAÇÃO CANÔNICA		FUNÇÃO APÓS	λ DE WILKS	QUI QUADRADO	GL	SIG.
					:	0	0,1664	44,831	10	0,00
1*	3,8190	93,93	93,93	0,8902	:	1	0,8020	5,517	4	0,24
2*	0,2469	6,07	100,00	0,4450	:					

* Assinala as duas funções discriminantes canônicas restantes na análise.

COEFICIENTES DA FUNÇÃO DISCRIMINANTE CANÔNICA PADRONIZADA

	FUNC 1	FUNC 2
RENDA	1,04740	– 0,42076
VIAGENS	0,33991	0,76851
FÉRIAS	– 0,14198	0,53354
TAMANHO DA CASA	– 0,16317	0,12932
IDADE	0,49474	0,52447

MATRIZ ESTRUTURAL

Correlações combinadas dentro de grupos entre variáveis discriminantes e funções discriminantes canônicas (variáveis ordenadas por tamanhos da correlação dentro da função)

	FUNC 1	FUNC 2
RENDA	0,85556*	– 0,27833
TAMANHO DA CASA	0,19319*	0,07749
FÉRIAS	0,21935	0,58829*
VIAGENS	0,14899	0,45362*
IDADE	0,16576	0,34079*

COEFICIENTES NÃO-PADRONIZADOS DA FUNÇÃO DISCRIMINANTE CANÔNICA

	FUNC 1	FUNC 2
RENDA	0,1542658	– 0,6197148E-01
VIAGENS	0,1867977	0,4223430
FÉRIAS	– 0,6952264E-01	0,2612652
TAMANHO DA CASA	– 0,1265334	0,1002796
IDADE	0,5928055E-01	0,6284206E-01
(constante)	– 11,09442	– 3,791600

(*continua*)

TABELA 18.5

Resultados da análise discriminante de três grupos *(Continuação)*

FUNÇÕES DISCRIMINANTES CANÔNICAS CALCULADAS EM MÉDIAS DE GRUPOS (CENTRÓIDES DE GRUPOS)

GRUPO	FUNC 1	FUNC 2
1	– 2,04100	0,41847
2	– 0,40479	– 0,65867
3	2,44578	0,24020

RESULTADOS DE CLASSIFICAÇÃO

			AFILIAÇÃO PREVISTA AO GRUPO			
		NÚMERO DE CASOS	1	2	3	TOTAL
Original	Contagem	1	9	1	0	10
		2	1	9	0	10
		3	0	2	8	10
	%	1	90,0	10,0	0,0	100,0
		2	10,0	90,0	0,0	100,0
		3	0,0	20,0	80,0	100,0
Validação cruzada	Contagem	1	7	3	0	10
		2	4	5	1	10
		3	0	2	8	10
	%	1	70,0	30,0	0,0	100,0
		2	40,0	50,0	10,0	100,0
		3	0,0	20,0	80,0	100,0

[a] A validação cruzada é feita somente para os casos que estão na análise. Na validação cruzada, cada caso é classificado pela função deduzida de todos os outros casos.
[b] 86,7% dos casos agrupados originais classificados corretamente.
[c] 66,7% dos casos agrupados de validação cruzada classificados corretamente.

RESULTADOS DE CLASSIFICAÇÃO PARA CASOS NÃO-SELECIONADOS PARA USO NA ANÁLISE

	GRUPO EFETIVO	NÚMERO DE CASOS	AFILIAÇÃO PREVISTA AO GRUPO 1	2	3
Grupo	1	4	3	1	0
			75,0%	25,0%	0,0%
Grupo	2	4	0	3	1
			0,0%	75,0%	25,0%
Grupo	3	4	1	0	3
			25,0%	0,0%	75,0%

Percentagem de casos agrupados classificados corretamente: 75,00%.

A Figura 18.2 é um gráfico em forma de diagrama de dispersão de todos os grupos sobre a função 1 e a função 2. Pode-se ver que o grupo 3 tem o valor máximo na função 1, e o grupo 1 tem o valor mínimo. Como a função 1 está associada essencialmente à renda e ao tamanho da casa, seria de esperar que os três grupos se apresentassem ordenados em relação a essas duas variáveis. Aqueles com maiores rendas e maior tamanho de casa tendem a gastar mais dinheiro com férias. Inversamente, aqueles com baixa renda e pequenas casas tendem a gastar menos em férias. Essa interpretação é reforçada ainda por um exame das médias de grupo em termos de renda e tamanho da casa.

A Figura 18.2 indica ainda que a função 2 tende a separar o grupo 1 (maior valor) e o grupo 2 (menor valor). Essa função está associada essencialmente às viagens, às férias e à idade. Dadas as correlações positivas dessas variáveis com a função 2 na matriz estrutural, é de esperar que o grupo 1 seja mais alto que o grupo 2 em termos de viagens, férias e idade. Isso, na verdade, é válido para viagens e férias, conforme mostram as médias de grupo dessas variáveis. Se as famílias no grupo 1 têm atitude mais favorável em relação às viagens e atribuem maior importância às férias do que as do grupo 2, por que gastam menos? Provavelmente gastariam mais com férias, se pudessem, mas isso é impedido pelas baixas rendas.

Obtém-se resultado análogo mediante exame de um ***mapa territorial***, conforme a Figura 18.3. Em um mapa territorial, cada centróide de grupo é indicado por um asterisco. As fronteiras do grupo são indicadas por números correspondentes aos grupos. Assim, o centróide do grupo 1 é limitado por 1s; o centróide do grupo 2, por 2s; e o centróide do grupo 3, por 3s.

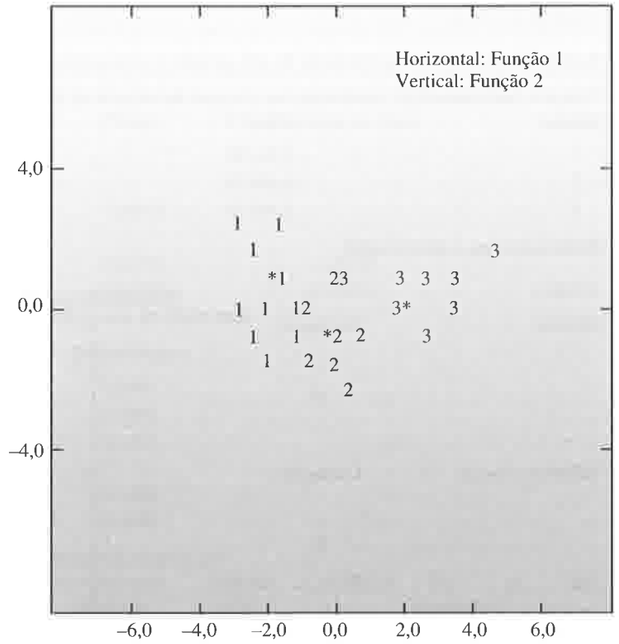

Figura 18.2 Diagrama de dispersão de todos os grupos.

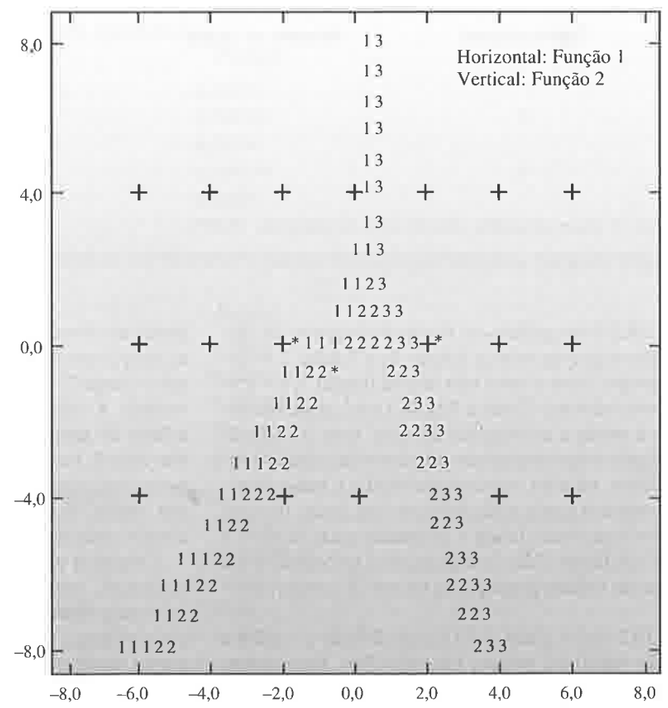

Figura 18.3 Mapa territorial.

mapa territorial
Recurso para avaliar os resultados da análise discriminante, representando em um gráfico a filiação ao grupo em cada caso.

AVALIAR A VALIDADE DA ANÁLISE DISCRIMINANTE

Os resultados de classificação baseados na amostra de análise indicam que (9 + 9 + 8)/30 = 86,67% dos casos estão classificados corretamente. A validação cruzada que exclui um elemento classifica corretamente apenas (7 + 5 + 8)/30 = 0,667 ou 66,7% dos casos. Quando a análise de classificação é feita na amostra retida independente da Tabela 18.3, obtém-se uma proporção de acerto de (3 + 3 + 3)/12 = 75%. Dados três grupos de igual tamanho, apenas pelo acaso poderíamos esperar uma proporção de acerto de 1/3 = 0,333, ou 33,3%. A melhora em relação ao acaso é de 50%, o que indica uma validade pelo menos satisfatória.

PESQUISA REAL

Lar é onde o coração do paciente está

Até 2003, o maior setor da economia dos Estados Unidos era o setor de serviços de saúde. Durante os sete anos decorrentes, até 2010, espera-se que os gastos com serviços de saúde cresçam significativamente mais rápido que a economia. Contribuem para a perspectiva positiva para esse segmento os dados demográficos atuais, especialmente com o aumento na demanda de cuidados de longo prazo à medida que a população envelhece. Espera-se que o número de estadunidenses que tem 85 anos ou mais triplique até o ano 2020 e, com esse grande aumento, é crucial que o sistema de saúde seja mostrado de forma positiva para esse segmento da população. Realizou-se uma pesquisa entre consumidores para determinar suas atitudes em relação a quatro sistemas de plano de saúde (atendimento a domicílio, hospitais, clínicas de repouso e ambulatórios) segundo 10 atributos. Obteve-se um total de 102 respostas e os resultados foram analisados com a análise discriminante múltipla (Tabela 1). Identificaram-se três funções discriminantes. Testes qui-quadrado sobre os resultados mostraram que as três funções discriminantes eram significativas ao nível de 0,01. A primeira função respondeu por 63% do poder discriminante total e as duas funções restantes contribuíram com 29,4 e 7,6%, respectivamente.

A Tabela 1 mostra os coeficientes da função discriminante padronizada das 10 variáveis das equações discriminantes. Os coeficientes variaram de –1 a +1. Para determinar a capacidade de cada atributo de classificar o sistema de saúde, foram utilizados valores absolutos. Na primeira função discriminante, as duas variáveis com os maiores coeficientes foram conforto (0,53) e privacidade (0,40). Como ambas dizem respeito à atenção e ao cuidado pessoal, a primeira dimensão foi rotulada "cuidado personalizado". Na segunda função, as duas variáveis com os maiores coeficientes foram qualidade do atendimento médico (0,67) e probabilidade de recuperação rápida (0,32). Assim, essa dimensão foi rotulada como "qualidade do atendimento médico". Na terceira função discriminante, os atributos mais significativos foram saneamento (–0,70) e despesa (0,52).

A análise discriminante múltipla mostra que as clínicas de repouso são encaradas desfavoravelmente quanto aos aspectos de cuidado pessoal, qualidade do atendimento médico e valor.

Como esses dois atributos representam valor e preço, a terceira função discriminante foi rotulada "valor".

A Tabela 2 exibe os quatro centróides de grupo. Essa tabela mostra que o atendimento médico a domicílio foi o atributo avaliado mais favoravelmente na dimensão do cuidado médico personalizado; os hospitais foram o elemento que teve avaliação menos favorável. Na dimensão da qualidade do atendimento médico, verificou-se substancial separação entre clínicas médicas e os outros três sistemas. Além disso, o atendimento médico a domicílio teve avaliações mais elevadas quanto à qualidade do atendimento médico que os ambulatórios. Por outro lado, os ambulatórios foram considerados os que oferecem melhor valor.

A análise de classificação das 102 respostas, detalhada na Tabela 3, acusou classificações corretas variando de 86% para hospitais a 68% para ambulatórios. A classificação incorreta para hospitais foi de 6% tanto para clínicas médicas como para ambulatórios e de 2% para atendimento médico a domicílio. As clínicas de repouso apresentaram classificação incorreta de 9% para hospitais, 10% para ambulatórios e 3% para atendimento médico a domicílio. Para ambulatórios, 9% das classificações incorretas corresponderam a hospitais, 13% a clínicas de repouso e 10% a atendimento médico a domicílio. Para este último, as classificações incorretas foram 5% para hospitais, 4% para clínicas de repouso e 13% para ambulatórios. Os resultados mostraram que as funções discriminantes foram suficientemente precisas para prever a filiação a um grupo.[12]

[12] Jan Tudor, "Valuation of the Health Services Industry," *Weekly Corporate Growth Report,* (1133) (March 26, 2001): 11237-11238; Kathryn H. Dansky e Diane Brannon, "Discriminant Analysis: A Technique for Adding Value to Patient Satisfaction Surveys," *Hospital & Health Services Administration,* 41 (4) (Winter 1996); 503-513; e Jeen-Su Lim e Ron Zallocco, "Determinant Attributes in Formulation of Attitudes Toward Four Health Care Systems," *Journal of Health Care Marketing* (June 1988): 25-30.

TABELA 1 – COEFICIENTES DA FUNÇÃO DISCRIMINANTE PADRONIZADA

	Função discriminante		
Produto	***1***	***2***	***3***
Seguro	– 0,20	– 0,04	0,15
Conveniente	0,08	0,08	0,07
Possibilidade de complicações médicas[a]	– 0,27	0,10	0,16
Dispendioso[a]	0,30	– 0,28	0,52
Confortável	0,53	0,27	– 0,19
Higiênico	– 0,27	– 0,14	– 0,70
Melhor cuidado médico	– 0,25	0,67	– 0,10
Privacidade	0,40	0,08	0,49
Rápida recuperação	0,30	0,32	– 0,15
Melhor pessoal médico	– 0,17	– 0,03	0,18
Percentagem explicada da variância	63,0	29,4	7,6
Qui-quadrado	663,3[b]	289,2[b]	70,1[b]

[a] Esses dois itens foram formulados negativamente no questionário. Foram codificados de maneira invertida para fins de análise de dados.
[b] $p < 0,01$.

TABELA 2 – CENTRÓIDES DOS SISTEMAS DE SAÚDE NO ESPAÇO DISCRIMINANTE

	Função discriminante		
Sistema	***1***	***2***	***3***
Hospital	– 1,66	0,97	– 0,08
Clínicas de repouso	– 0,60	– 1,36	– 0,27
Ambulatório	0,54	– 0,13	0,77
Assistência médica a domicílio	1,77	0,50	–0,39

TABELA 3 – TABELA DE CLASSIFICAÇÃO

	Classificação (%)			
Sistema	***Hospital***	***Clínicas de repouso***	***Ambulatório***	***Assistência médica a domicílio***
Hospital	86	6	6	2
Clínicas de repouso	9	78	10	3
Ambulatório	9	13	68	10
Assistência médica a domicílio	5	4	13	78 ■

ANÁLISE DISCRIMINANTE PASSO A PASSO

A análise discriminante passo a passo é análoga à regressão múltipla passo a passo (ver Capítulo 17) quanto a os previsores serem introduzidos seqüencialmente com base em sua capacidade de discriminar entre grupos. Calcula-se uma razão F para cada previsor, fazendo-se uma análise da variância univariada na qual os grupos são tratados como variável categórica e o previsor como variável critério. O previsor com maior razão F é o primeiro a ser selecionado para inclusão na função discriminante, desde que satisfaça certos critérios de significância e tolerância. Acrescenta-se um segundo previsor com base na maior razão F ajustada ou parcial, levando-se em conta o previsor já selecionado.

Cada previsor selecionado é testado quanto à retenção com base em sua associação com outros previsores selecionados. Continua-se o procedimento de seleção até que tenham sido introduzidos na função discriminante todos os previsores que satisfaçam os critérios de inclusão e retenção. Diversas estatísticas são calculadas em cada passo. Além disso, na conclusão, apresenta-se um resumo dos previsores introduzidos ou removidos. O procedimento passo a passo também oferece a saída padrão apresentada no método direto.

A opção pelo procedimento passo a passo se baseia no critério de otimização adotado. O ***procedimento de Mahalanobis*** se baseia na maximização de uma medida generalizada da distância entre os dois grupos mais próximos. Esse procedimento

permite aos pesquisadores de marketing utilizar ao máximo a informação disponível.[13]

procedimento de Mahalanobis
Procedimento passo a passo usado na análise discriminante para maximizar uma medida generalizada de distância entre os dois grupos mais próximos.

O método Mahalanobis foi utilizado para se fazer uma análise discriminante passo a passo de dois grupos sobre os dados relativos à variável "visita" nas Tabelas 18.2 e 18.3. A primeira variável previsora a ser selecionada foi a renda, seguida pelo tamanho da residência e pelas férias. A ordem em que as variáveis foram selecionadas também indica sua importância para fazer discriminação entre grupos – o que é corroborado por um exame dos coeficientes da função discriminante padronizada e pelos coeficientes de correlação estrutural. Observe que os resultados da análise passo a passo convergem para as conclusões referidas anteriormente pelo método direto.

O exemplo a seguir mostra uma aplicação da análise discriminante na pesquisa de marketing internacional, e o exemplo posterior apresenta uma aplicação com relação à ética.

PESQUISA REAL

Resultados satisfatórios de programas de satisfação na Europa

Atualmente, é cada vez maior o número de empresas de informática que procura se esmerar em seus serviços aos clientes, em vez de dar ênfase à capacidade e às características dos próprios computadores. A Hewlett-Packard aprendeu essa lição ao fazer negócios na Europa. A pesquisa feita no mercado europeu revelou uma diferença na ênfase dada às necessidades de serviço ao longo dos segmentos etários. Grupos de foco revelaram que os clientes com mais de 40 anos tinham dificuldades com aspectos técnicos do computador e dependiam em grande parte de programas de serviço ao consumidor. Por outro lado, os consumidores jovens apreciavam os aspectos técnicos do produto, que contribuíam para sua satisfação. Uma investigação posterior, na forma de uma grande pesquisa transversal, revelou fatores que conduziam a diferenças nos dois segmentos. Fez-se uma análise discriminante de dois grupos – clientes satisfeitos e clientes insatisfeitos – e com diversas variáveis independentes, tais como informação técnica, facilidade de operação, variedade e escopo dos programas de atendimento ao consumidor. Os resultados confirmaram que a diversidade e a abrangência desses programas constituía, na verdade, um importante fator de diferenciação. Foi uma constatação crucial, porque então a Hewlett-Packard passou a tratar melhor os clientes insatisfeitos, focalizando mais os serviços de atendimento do que os detalhes técnicos. Em conseqüência disso, a Hewlett-Packard lançou três programas bem-sucedidos de atendimento ao cliente – retorno do consumidor, pesquisas sobre a satisfação do cliente e controle da qualidade total. Esses programas resultaram em um aumento do grau de satisfação dos clientes. Depois de constatar o sucesso dos resultados desses programas na Europa, a HP Malaysia decidiu lançar um programa de serviços ao cliente chamado de HP Cares! em 2001. Esse programa foi desenvolvido na Malásia como parte da iniciativa Experiência Total do Consumidor (TCE – Total Customer Experience) da HP e está sendo implementado em vários pontos do mundo em 2004. O gerente de vendas e de marketing do HP Malaysia acredita que esse programa permitirá que a empresa construa valor de marca e diferencie a empresa de seus concorrentes. A HP percebe que não importa quão bons sejam seus produtos, a fidelidade à marca só pode ser alcançada através de um serviço contínuo ao cliente.[14] ■

PESQUISA REAL

Análise discriminante discrimina as firmas éticas e as antiéticas

Utilizou-se a análise discriminante para identificar as variáveis relevantes que predizem um comportamento ético, ou antiético. Uma pesquisa anterior sugeriu que as variáveis com impacto sobre as decisões éticas são atitudes, liderança, presença ou ausência de códigos de ética de conduta e o tamanho da organização.

Para determinar quais dessas variáveis são os melhores previsores do comportamento ético, pesquisaram-se 149 firmas, pedindo que indicassem como operavam em 18 situações éticas diferentes. Dessas 18 situações, nove se referiam a atividades de marketing, que incluíam propaganda enganosa, aceitação de presentes em troca de tratamento preferencial, preço abaixo do custo e semelhantes. Com base nessas nove questões, as firmas entrevistadas foram classificadas em dois grupos: "as que nunca praticam" e "as que praticam".

Um exame das variáveis que influenciaram a classificação através da análise discriminante de dois grupos mostrou que as atitudes e o tamanho da firma eram os melhores previsores do comportamento ético. Evidentemente, as firmas menores tendem a apresentar comportamento mais ético em problemas de marketing. Uma empresa em particular que se concentra em realizar práticas éticas de negócios é o Smile Internet Bank no Reino Unido. No início de 2002, o grupo de marketing Smile lançou seis personagens de desenho que se centram na posição ética do banco. Cada personagem simboliza um dos seis traços bancários negativos e finalmente apresenta a Smile como empresa que oferece o oposto desses traços. Essa estratégia de marketing foi muito bem-sucedida.[15] ■

[13] Richard A. Johnson e Dean W. Wichern, *Applied Multivariate Statistical Analysis,* 5th ed. (Upper Saddle River, NJ: Prentice-Hall, 2002); Joseph F. Hair, Jr., Ralph E. Anderson, Ronald L. Tatham e William C. Black. *Multivariate Data Analysis with Readings,* 5th ed. (Englewood Cliffs, NJ: Prentice-Hall, Inc., 1998).

[14] Prasanna Raman, "Taking Customer Service a Step Ahead," *Computimes* (October 22, 2001): 1; e Charlotte Klopp e John Sterlicchi, "Customer Satisfaction Just Catching on in Europe," *Marketing News* (May 28, 1990).

[15] Suzy Bashford, "Smile in Ethics Push with Cartoon Icons," *Marketing* (January 10, 2002): 1; e Paul R. Murphy, Jonathan E. Smith e James M. Daley, "Executive Attitudes, Organizational Size and Ethical Issues: Perspectives on a Service Industry," *Journal of Business Ethics,* 11 (1992): 11-19.

APLICAÇÕES NA INTERNET E EM COMPUTADORES

Na versão do SPSS para *mainframe*, o procedimento DISCRIMINANT é usado para a realização de análise discriminante. Trata-se de um programa geral que pode ser usado para análise discriminante de dois grupos ou múltipla. Além disso, o método direto ou passo a passo pode ser adotado.

No SAS, o procedimento DSCRIM pode ser usado para a realização de análise discriminante de dois grupos ou múltipla. Se não for possível atingir a suposição de uma distribuição multivariada normal, o procedimento NEIGHBOR pode ser usado. Nesse procedimento, uma regra não-paramétrica do vizinho mais próximo é usada para classificar as observações. CANDISC desempenha a análise canônica discriminante e é relacionado à análise de componente principal e à correlação canônica. O procedimento STEPDISC pode ser usado para realizar a análise discriminante passo a passo. As versões para *mainframe* e PCs são similares, a não ser pelo programa NEIGHBOR, que não está disponível na versão para microcomputador.

No MINITAB, pode-se fazer a análise discriminante utilizando-se a função Stat>Multivariate>Discriminate Analysis, que calcula a análise discriminante tanto linear como quadrática na classificação das observações em dois ou mais grupos. A análise discriminante não está incluída no EXCEL.

SPSS Windows

O programa DISCRIMINANT desempenha tanto a análise discriminante de dois grupos quanto a múltipla. Para escolher esse procedimento usando o SPSS para Windows, clique em:

Analyze>Classify>Discriminant...

FOCO NA BURKE

Tal como no caso da regressão múltipla, as principais aplicações da análise discriminante são a previsão e a determinação da importância relativa das variáveis previsoras. A diferença chave entre essas duas técnicas é a exigência, na regressão múltipla, de uma variável dependente intervalar ou razão, ao passo que a análise discriminante utiliza uma variável dependente dicotômica ou categórica. Enquanto a regressão múltipla pode ser usada para prever o grau de interesse por uma compra, a análise discriminante seria utilizada somente se quiséssemos prognosticar a formação de grupos de compradores e não-compradores.

Um dos problemas que os pesquisadores devem enfrentar é a utilização de "grupos naturais" *versus* "grupos criados". Por exemplo, ao pesquisarmos a idade de um grupo de entrevistados, em geral coletamos os dados em categorias, e não no número exato de anos. Ao fazermos isso, as categorias utilizadas determinam o resultado de uma análise em que elas constituem a variável dependente. Misturando as pessoas em grupos etários arbitrários, podemos disfarçar as diferenças entre essas pessoas. Considerando os dados de forma mais granular, podemos formular julgamentos sobre a criação de grupos maiores de entrevistados com base em qualquer medida em particular. Em outra situação, podemos também colher dados sobre se a pessoa vive em uma casa, em um apartamento ou em um condomínio. Como as categorias nominais são relativamente "naturais", seria difícil achar categorias que descrevem a residência que tivessem qualidade mais intervalar. A confiança no significado dos grupos permite confiarmos na validade das conclusões.

Em um estudo da Burke, a análise da utilização de um produto resultou na seguinte matriz de classificação. Cada entrevistado foi consultado sobre se havia experimentado determinado produto. Procurou-se determinar a discriminação com base nas avaliações dos entrevistados sobre os benefícios esperados de um produto dessa categoria. Todo entrevistado foi considerado como um usuário da categoria. Na amostra, 30% (60) não tinham experimentado o produto e 70% (202) já o tinham experimentado. Em geral, quando os grupos são desproporcionais na população, espera-se que a previsão referente ao grupo maior seja mais precisa que a do grupo menor. Uma forma de encarar esse fato é que, mediante escolha aleatória, classificaríamos corretamente um usuário de uma marca 59,5% das vezes (0,771 × 0,771) e um não-usuário, 5,2% das vezes (0,229 × 0,229), e as pessoas seriam classificadas incorretamente em 35,4% das vezes (0,229 × 0,771 + 0,771 × 0,229). Esse modelo classifica o grupo menor (não usuários da marca) com menor precisão que os usuários, mas, mesmo assim, deu melhor resultado do que se poderia esperar de uma escolha aleatória.

Utilizou-se a opção de validação cruzada em SPSS. Nessa situação, a validação cruzada produziu quase o mesmo nível de precisão que o modelo total. Observe que a validação cruzada gerou mais erros ao fazer previsões para o grupo menor.

O resultado típico nem sempre se presta às alternativas que a administração possa desejar para interpretar o modelo à luz da utilização pretendida. Por exemplo, a direção pergunta, "se o leitor fosse prever que um entrevistado é um não-usuário, qual a probabilidade estimada de que o leitor esteja certo?" Em tal situação (com os resultados da validação cruzada), a Burke diria, "provavelmente preveríamos que um entrevistado é não-usuário em cerca de 18% das vezes [(33 + 15)/262 = 0,18]. Assim fazendo, esperaríamos estar certos 69% das vezes (33/48 = 0,69)". O resultado padrão não se apresenta de forma a permitir que a direção veja imediatamente essa resposta. Novamente referindo-se à possibilidade, Burke diria à gerência, "a seleção aleatória estimaria que cerca de 23(22,9)% seriam não-usuários e estaria correta apenas 23(22,9)% das vezes em que isso ocorre" (isto é, a proporção de uma amostra estimada aleatoriamente como não usuários é 0,229 × 0,771 + 0,229 × 0,229 = 0,229; proporção prevista corretamente = 0,229 × 0,229/(0,229 × 0,229 + 0,229 0,771) = 22,9%). Considerar apenas a percentagem classificada corretamente não nos diz tudo.

RESULTADOS DA CLASSIFICAÇÃO[b,c]

			Afiliação prevista ao grupo		
		Experimentou a marca	*Não*	*Sim*	*Total*
Original	Contagem	Não	35	25	60
		Sim	15	187	202
	%	Não	58,3	41,7	100,0
		Sim	7,4	92,6	100,0
Validação cruzada[a]	Contagem	Não	33	27	60
	%	Não	55,0	45,0	100,0
		Sim	7,4	92,6	100,0

[a] A validação cruzada é feita somente para os casos que estão na análise. Na validação cruzada, cada caso é classificado pela função deduzida de todos os outros casos.
[b] 84,7% dos casos originais grupados classificados corretamente.
[c] 84,0% dos casos grupados em validação cruzada classificados corretamente.

(*continua*)

RESUMO

A análise discriminante é útil para analisar dados quando a variável critério, ou dependente, é categórica e as variáveis independentes, ou previsoras, são escalonadas por intervalo. Quando a variável critério tem duas categorias, a técnica é conhecida como análise discriminante de dois grupos. A análise discriminante múltipla diz respeito ao caso em que estão em jogo três ou mais categorias.

O procedimento de análise discriminante consiste em cinco passos. Em primeiro lugar, a formulação do problema discriminante requer a identificação dos objetivos e as variáveis de critério (dependente) e previsoras. A amostra é dividida em duas partes. Uma parte, a amostra de análise, serve para estimar a função discriminante. A outra, a amostra retida, é reservada para a validação. A estimação – o segundo passo – exige a formulação de uma combinação linear dos previsores, chamados de funções discriminantes, de modo que os grupos difiram tanto quanto possível nos valores previsores.

A determinação da significância estatística é o terceiro passo. Envolve o teste da hipótese nula de que, na população, as médias de todas as funções discriminantes em todos os grupos sejam iguais. Se a hipótese nula for rejeitada, faz sentido interpretar os resultados.

O quarto passo, a interpretação dos pesos ou coeficientes discriminantes, é análogo ao da análise de regressão múltipla. Dada a multicolinearidade nas variáveis previsoras, não há medida não-ambígua da importância relativa dos previsores na discriminação entre os grupos. Todavia, pode-se ter alguma idéia da importância relativa das variáveis examinando-se a magnitude absoluta dos coeficientes da função discriminante padronizada e as correlações estruturais ou cargas discriminantes. Essas correlações simples entre cada previsor e a função discriminante representam a variância que o previsor compartilha com a função. Outra ajuda na interpretação dos resultados da análise discriminante consiste em estabelecer um perfil característico para cada grupo, com base nas médias de grupo para as variáveis previsoras.

A validação – o quinto passo – estabelece a matriz de classificação. Os pesos discriminantes, estimados com auxílio da amostra de análise, são multiplicados pelos valores das variáveis previsoras na amostra retida para gerar escores discriminantes para os casos naquela amostra. Os casos são, então, atribuídos a grupos, com base em seus escores discriminantes e em uma regra de decisão apropriada. Determina-se a percentagem de casos classificados corretamente e compara-se com a percentagem esperada de uma classificação aleatória.

Há duas amplas abordagens para estimar os coeficientes. O método direto envolve a estimação da função discriminante, de forma que todos os previsores sejam incluídos simultaneamente. Uma alternativa é o método passo a passo, no qual as variáveis previsoras são introduzidas seqüencialmente, de acordo com sua capacidade de discriminar entre grupos.

Na análise discriminante múltipla, se há G grupos e k previsores, é possível estimar até $G - 1$, ou k (o que for menor) funções discriminantes. A primeira função tem a maior razão de somas de quadrados entre grupos para soma de quadrados dentro de grupos. A segunda função, não correlacionada com a primeira, tem a segunda maior razão, e assim por diante.

TERMOS E CONCEITOS FUNDAMENTAIS

análise discriminante, 524
função discriminante, 524
análise discriminante de dois grupos, 524
análise discriminante múltipla, 524
modelo de análise discriminante, 525
correlação canônica, 525
centróide, 525
matriz de classificação, 525
coeficientes da função discriminante, 526
escores discriminantes, 526
autovalores, 526
valores F e sua significância, 526
médias de grupo e desvios-padrão de grupo, 526
matriz de correlação combinada dentro de grupos, 526
coeficientes da função discriminante padronizada, 526
correlações estruturais, 526
matriz de correlação total, 526
λ de Wilks, 526
amostra de análise, 526
amostra de validação, 526
método direto, 531
análise discriminante passo a passo, 531
perfil das características, 533
proporção de acertos, 533
mapa territorial, 539
procedimento de Mahalanobis, 541

EXERCÍCIOS

Perguntas

1. Quais são os objetivos da análise discriminante?
2. Qual é a principal distinção entre análise discriminante de dois grupos e múltipla?
3. Descreva a relação da análise discriminante com regressão e ANOVA.
4. Quais são as etapas envolvidas no procedimento da análise discriminante?
5. Como se deve dividir a amostra total para fins de estimação e validação?
6. O que é o λ de Wilks e para que é utilizado?
7. Defina escores discriminantes.
8. Explique o que significa um autovalor.
9. O que é matriz de classificação?
10. Explique o conceito de correlações estruturais.
11. Como se determina a significância estatística da análise discriminante?
12. Descreva um procedimento usual para determinar a validade da análise discriminante.
13. Quando os grupos são todos de mesmo tamanho, como se determina a precisão da classificação aleatória?
14. Em que diferem o procedimento discriminante passo a passo e o método direto?

Problemas

1. Em uma investigação das diferenças entre usuários freqüentes e usuários eventuais de alimentos congelados, constatou-se que os dois maiores coeficientes da função discriminante padronizada eram 0,97 para orientação de conveniência e 0,61 para renda. É correto concluir que a orientação de conveniência é mais importante do que a renda quando cada variável é considerada por si mesma?
2. Dadas as informações abaixo, calcule o escore discriminante para cada entrevistado. O valor da constante é 2,04.

Coeficientes da função discriminante não-padronizada	
Idade	0,38
Renda	0,44
Risco assumido	−0,39
Otimista	1,26

ID do entrevistado	*Idade*	*Renda*	*Risco assumido*	*Otimista*
0246	36	43,7	21	65
1337	44	62,5	28	56
2375	57	33,5	25	40
2454	63	38,7	16	36

EXERCÍCIOS PARA INTERNET E COMPUTADORES

1. Faça uma análise discriminante de dois grupos para os dados das Tabelas 18.2 e 18.3 utilizando pacotes SPSS, SAS e MINITAB (microcomputador ou *mainframe*). Compare os resultados de todos os pacotes. Discuta as semelhanças e as diferenças.
2. Faça uma análise discriminante de três grupos passo a passo sobre os dados das Tabelas 18.2 e 18,3 utilizando um pacote SPSS, SAS ou MINITAB. Compare os resultados com os da Tabela 18.5 para análise discriminante de três grupos.
3. Analise os dados sobre a Nike apresentados no exercício para Internet e computadores 2, no Capítulo 15. Os três grupos de usos diferem em termos de consciência, atitude, preferência, intenção e fidelidade à marca quando essas variáveis são consideradas simultaneamente?
4. Analise os dados sobre estilo de vida fora de casa apresentados no exercício para Internet e computadores 2 do Capítulo 15. Os três grupos baseados em local de residência diferem sobre a importância atribuída a desfrutar da natureza, relacionar-se com o clima, viver em harmonia com o ambiente, exercitar-se regularmente e encontrar-se com outras pessoas (V_2 a V_6) quando essas variáveis são consideradas simultaneamente?
5. Faça uma análise discriminante de dois grupos sobre os dados que você obteve no exercício de Trabalho de Campo 1, usando o pacote SPSS, SAS ou MINITAB. É possível diferenciar alunos com graduação de alunos com pós-graduação usando as quatro medidas de atitude?

ATIVIDADES

Trabalho de campo

1. Entreviste 15 alunos de graduação e 15 alunos de pós-graduação. Meça suas atitudes com relação ao ensino universitário (se valeu a pena obter um diploma universitário), a desfrutar da vida (se é importante se divertir na vida), a sua universidade (não estão muito felizes por ter escolhido estudar aqui) e a ética no trabalho (geralmente, há falta de ética no trabalho no *campus* da faculdade). Para cada atitude, meça o grau de discordância/concordância usando uma escala de classificação de sete pontos (1 = discordo, 7 = concordo).

Discussão em grupo

2. É significativo determinar a importância relativa dos previsores ao discriminar grupos? Por quê, ou por que não? Discuta em um pequeno grupo.

APÊNDICE 18.1

Estimação dos coeficientes da função discriminante

Suponha que haja G grupos, $i = 1, 2, 3, \ldots, G$, cada um contendo n_i observações sobre K variáveis independentes, $X_1, X_2 \ldots, X_k$. Emprega-se a notação seguinte:

N = tamanho total da amostra

$$= \sum_{i=1}^{G} n_i$$

$\boldsymbol{W}_i$ = matriz da média corrigida de soma de quadrados e produtos cruzados para o i-ésimo grupo

$\boldsymbol{W}$ = matriz da média corrigida de soma de quadrados, combinada, dentro de grupos e produtos cruzados

$\boldsymbol{B}$ = matriz da média corrigida de soma de quadrados entre grupos e produtos cruzados

$\boldsymbol{T}$ = matriz da média total, corrigida, de soma de quadrados e produtos cruzados para todas as N observações ($= \boldsymbol{W} + \boldsymbol{B}$)

$\bar{\boldsymbol{X}}_i$ = vetor de média de observações no i-ésimo grupo

$\bar{\boldsymbol{X}}$ = vetor das médias globais para todas as N observações

λ = razão das somas de quadrados entre grupos para as somas de quadrados dentro de grupos

$\boldsymbol{b}$ = vetor dos coeficientes discriminantes ou pesos

Então,

$$\boldsymbol{T} = \sum_{i=1}^{G}\sum_{j=1}^{n_i}\left(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}\right)\left(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}\right)'$$

$$\boldsymbol{W}_i = \sum_{j=1}^{n_i}\left(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i\right)\left(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i\right)'$$

$$\boldsymbol{W} = \boldsymbol{W}_1 + \boldsymbol{W}_2 + \boldsymbol{W}_3 + \cdots + \boldsymbol{W}_G$$

$$\boldsymbol{B} = \boldsymbol{T} - \boldsymbol{W}$$

Definamos o composto linear $\boldsymbol{D} = \boldsymbol{b}'_1\boldsymbol{X}$. Então, com referência a $\boldsymbol{D}$, as somas de quadrados entre grupos e dentro de grupos são, respectivamente, dadas por $\boldsymbol{b}'_1\boldsymbol{B}\boldsymbol{b}$ e $\boldsymbol{b}'_1\boldsymbol{W}\boldsymbol{b}$. Para maximizar a discriminação dos grupos, estimam-se as funções discriminantes de modo a maximizarem a variabilidade entre grupos. Calculam-se os coeficientes $\boldsymbol{b}$ que maximizam λ, resolvendo

$$\text{Max } \lambda = \frac{\boldsymbol{b}'\boldsymbol{B}\boldsymbol{b}}{\boldsymbol{b}'\boldsymbol{W}\boldsymbol{b}}$$

Tomando a derivada parcial em relação a λ e igualando-a a zero, obtemos, após algumas simplificações:

$$(\boldsymbol{B} - \lambda\boldsymbol{W})\,\boldsymbol{b} = 0$$

Para resolver em relação a b, é mais conveniente pré-multiplicar por W^{-1} e resolver a seguinte equação característica:

$$(\boldsymbol{W}^{-1}\boldsymbol{B} - \lambda\boldsymbol{I})\,\boldsymbol{b} = 0$$

O valor máximo de λ é o maior autovalor da matriz $\boldsymbol{W}^{-1}\boldsymbol{B}$, e $\boldsymbol{b}$ é o autovetor associado. Os coeficientes de b são os coeficientes discriminantes, ou pesos, associados à primeira função discriminante. Em geral é possível estimar até $G - 1$ ou k (o que for menor) funções discriminantes, cada uma com seu autovalor associado. As funções discriminantes são estimadas seqüencialmente. Em outras palavras, a primeira função discriminante esgota a maior parte da variabilidade entre grupos. A segunda função maximiza a variação entre grupos que não foi explicada pela primeira, e assim por diante.