CAPÍTULO 5

Análise Discriminante Múltipla e Regressão Logística

Objetivos de aprendizagem

Ao concluir este capítulo, você deverá ser capaz de:

- Estabelecer as circunstâncias sob as quais a análise discriminante linear ou a regressão logística deve ser usada no lugar de uma regressão múltipla.
- Identificar as questões mais importantes relativas aos tipos de variáveis usadas e ao tamanho de amostra exigido na aplicação de análise discriminante.
- Compreender as suposições inerentes à análise discriminante para avaliar a adequação de seu uso em um problema em particular.
- Descrever as duas abordagens computacionais para a análise discriminante e o método para avaliar o ajuste geral do modelo.
- Explicar o que é uma matriz de classificação e como desenvolver uma, e descrever as maneiras de avaliar a precisão preditiva da função discriminante.
- Dizer como identificar variáveis independentes com poder discriminatório.
- Justificar o uso de uma abordagem de partição de amostras para validação.
- Compreender as vantagens e desvantagens da regressão logística comparada com a análise discriminante e a regressão múltipla.
- Interpretar os resultados de uma análise de regressão logística, comparando-os com a regressão múltipla e a análise discriminante.

Apresentação do capítulo

A regressão múltipla é sem dúvida a técnica de dependência multivariada mais amplamente empregada. A base para a popularidade da regressão tem sido sua habilidade de prever e explicar variáveis métricas. Mas o que acontece quando variáveis não-métricas tornam a regressão múltipla inadequada? Este capítulo introduz duas técnicas – análise discriminante e regressão logística – que tratam da situação de uma variável dependente não-métrica. Neste tipo de situação, o pesquisador está interessado na previsão e na explicação das relações que afetam a categoria na qual um objeto está localizado, como a questão do por quê uma pessoa é um cliente ou não, ou se uma empresa terá sucesso ou fracassará. Os dois maiores objetivos deste capítulo são:

1. Introduzir a natureza, a filosofia e as condições da análise discriminante múltipla e da regressão logística
2. Demonstrar a aplicação e interpretação dessas técnicas com um exemplo ilustrativo

O Capítulo 1 estabeleceu que o propósito básico da análise discriminante é estimar a relação entre uma variável dependente não-métrica (categórica) e um conjunto de variáveis independentes métricas, nesta forma geral:

$$Y_1 = X_1 + X_2 + X_3 + \cdots + X_n$$

(não-métrica) (métricas)

A análise discriminante múltipla e a regressão logística encontram amplas aplicações em situações nas quais o objetivo principal é identificar o grupo ao qual um objeto (p.ex., uma pessoa, uma firma ou um produto) pertence. Aplicações potenciais incluem prever o sucesso ou fracasso de um novo produto, decidir se um estudante deve ser aceito em uma faculdade, classificar estudantes quanto a interesses vocacionais, determinar a categoria de risco de crédito de uma pessoa, ou prever se uma empresa terá sucesso. Em cada caso, os objetos recaem em grupos, e o objetivo é prever ou explicar as bases para a pertinência de cada objeto a um grupo através de um conjunto de variáveis independentes selecionadas pelo pesquisador

Termos-chave

Antes de começar o capítulo, leia os termos-chave para compreender os conceitos e a terminologia empregados. Ao longo do capítulo, os termos-chave aparecem em **negrito**. Outros pontos que merecem destaque, além das referências cruzadas nos termos-chave, estão em *itálico*. Exemplos ilustrativos estão em quadros.

Amostra de análise Grupo de casos usado para estimar a(s) *função(ões) discriminante(s)* ou o modelo de *regressão logística*. Quando se constroem *matrizes de classificação*, a amostra original é dividida aleatoriamente em dois grupos, um para estimação do modelo (a amostra de análise) e o outro para validação (a *amostra de teste*).

Abordagem de extremos polares Método para construir uma variável dependente categórica a partir de uma *variável métrica*. Primeiro, a variável métrica é dividida em três categorias. Em seguida, as categorias extremas são usadas na análise discriminante ou na regressão *logística*, e a categoria do meio não é incluída na análise.

Amostra de teste Grupo de objetos não usados para computar a(s) função(ões) discriminante(s) ou o modelo de *regressão logística*. Esse grupo é então usado para validar a função discriminante ou o modelo de regressão logística em uma amostra separada de respondentes. É também chamada de *amostra de validação*.

Amostra de validação Ver *amostra de teste*.

Análise *logit* Ver *regressão logística*.

Cargas discriminantes Medida da correlação linear simples entre cada variável independente e o *escore Z discriminante* para cada função discriminante; também chamadas de *correlações estruturais*. As cargas discriminantes são calculadas sendo incluída uma variável independente na função discriminante ou não.

Centróide Valor médio para os *escores Z discriminantes* de todos os objetos, em uma dada categoria ou grupo. Por exemplo, uma análise discriminante de dois grupos tem dois centróides, um para os objetos em cada grupo.

Coeficiente discriminante Ver *peso discriminante*.

Coeficiente logístico exponenciado Anti-logaritmo do *coeficiente logístico*, usado para fins de interpretação na regressão logística. O coeficiente exponenciado menos 1,0 é igual à mudança percentual nas desigualdades. Por exemplo, um coeficiente exponenciado de 0,20 representa uma mudança negativa de 80% na desigualdade (0,20 – 1,0 = – 0,80) para cada unidade de variação na variável independente (o mesmo se a desigualdade fosse multiplicada por 0,20). Assim, um valor de 1,0 se iguala a nenhuma mudança na desigualdade, e valores acima de 1,0 representam aumentos na desigualdade prevista.

Coeficiente logístico Coeficiente no modelo de regressão logística que atua como o fator de ponderação para as variáveis independentes em relação a seu poder discriminatório. Semelhante a um peso de regressão ou um *coeficiente discriminante*.

Correlações estruturais Ver *cargas discriminantes*.

Critério das chances proporcionais Outro critério para avaliar a *razão de sucesso*, no qual a probabilidade média de classificação é calculada considerando-se todos os tamanhos de grupos.

Critério de chance máxima Medida de precisão preditiva na *matriz de classificação* que é calculada como o percentual de respondentes no maior grupo. A idéia é que a melhor escolha desinformada é classificar cada observação no maior grupo.

Curva logística Uma curva em S formada pela *transformação logit* que representa a probabilidade de um evento. A forma em S é não-linear porque a probabilidade de um evento deve se aproximar de 0 e 1, porém jamais sair destes limites. Assim, apesar de haver uma componente linear no meio do intervalo, à medida que as probabilidades se aproximam dos limites inferior e superior de probabilidade (0 e 1), elas devem se amenizar e ficar assintóticas nesses limites.

Escore de corte ótimo Valor de *escore Z discriminante* que melhor separa os grupos em cada função discriminante para fins de classificação.

Escore de corte Critério segundo o qual cada *escore Z discriminante* individual é comparado para determinar a pertinência prevista em um grupo. Quando a análise envolve dois grupos, a previsão de grupo é determinada computando-se um único escore de corte. Elementos com escores *Z* discriminantes abaixo dessa marca são designados a um grupo, enquanto aqueles com escores acima são classificados no outro. Para três ou mais grupos, funções discriminantes múltiplas são usadas, com um escore de corte diferente para cada função.

Escore Z Ver *escore Z discriminante*.

Escore *Z* discriminante Escore definido pela *função discriminante* para cada objeto na análise e geralmente dado em termos padronizados. Também conhecido como escore *Z*, é calculado para cada objeto em cada função discriminante e usado em conjunção com o *escore de corte* para determinar pertinência prevista ao grupo. É diferente da terminologia escore *z* usada para variáveis padronizadas.

Estatística *Q* de Press Medida do poder classificatório da *função discriminante* quando comparada com os resultados

esperados de um modelo de chances. O valor calculado é comparado com um valor crítico baseado na distribuição quiquadrado. Se o valor calculado exceder o valor crítico, os resultados da classificação serão significantemente melhores do que se esperaria do acaso.

Estatística Wald Teste usado em *regressão logística* para a significância do *coeficiente logístico*. Sua interpretação é semelhante aos valores *F* ou *t* usados para o teste de significância de coeficientes de regressão.

Estimação simultânea Estimação da(s) *função(ões) discriminante(s)* ou do modelo de *regressão logística* em um único passo, onde pesos para todas as variáveis independentes são calculados simultaneamente; contrasta com a *estimação stepwise*, na qual as variáveis independentes entram seqüencialmente de acordo com o poder discriminante.

Estimação *stepwise* Processo de estimação de *função(ões) discriminante(s)* ou do modelo de *regressão logística* no qual variáveis independentes entram seqüencialmente de acordo com o poder discriminatório que elas acrescentam à previsão de pertinência no grupo.

Expansão dos vetores *Vetor escalonado* no qual o vetor original é modificado para representar a razão *F* correspondente. Usado para representar graficamente as *cargas da função discriminante* de uma maneira combinada com os *centróides* de grupo.

Função de classificação Método de classificação no qual uma função linear é definida para cada grupo. A classificação é realizada calculando-se um escore para cada observação na função de classificação de cada grupo e então designando-se a observação ao grupo com o maior escore. É diferente do cálculo do *escore Z discriminante*, que é calculado para cada *função discriminante*.

Função discriminante linear de Fisher Ver *função de classificação*.

Função discriminante Uma variável estatística das variáveis independentes selecionadas por seu poder discriminatório usado na previsão de pertinência ao grupo. O valor previsto da função discriminante é o *escore Z discriminante*, o qual é calculado para cada objeto (pessoa, empresa ou produto) na análise. Ele toma a forma da equação linear

$$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + \cdots + W_nX_{nk}$$

onde

Z_{jk} = escore **Z** discriminante da função discriminante *j* para o objeto *k*
a = intercepto
W_i = peso discriminante para a variável independente *i*
X_{ik} = variável independente *i* para o objeto *k*

Índice potência Medida composta do poder discriminatório de uma variável independente quando mais de uma *função discriminante* é estimada. Baseada em *cargas discriminantes*, é uma medida relativa usada para comparar a discriminação geral dada por conta de cada variável independente em todas as funções discriminantes significantes.

***M* de Box** Teste estatístico para a igualdade das matrizes de covariância das variáveis independentes nos grupos da variável dependente. Se a significância estatística não exceder o nível crítico (i.e., não-significância), então a igualdade das matrizes de covariância encontra sustentação. Se o teste mostra significância estatística, os grupos são considerados diferentes e a suposição é violada.

Mapa territorial Representação gráfica dos escores de *corte* em um gráfico de duas dimensões. Quando é combinado com os gráficos de casos individuais, a dispersão de cada grupo pode ser vista e as classificações ruins de casos individuais podem ser diretamente identificadas a partir do mapa.

Matriz de classificação Meio de avaliar a habilidade preditiva da(s) função(ões) discriminante(s) ou da *regressão logística* (também chamada de matriz confusão, designação ou de previsão). Criada pela tabulação cruzada dos membros do grupo real com os do grupo previsto, essa matriz consiste em números na diagonal, que representam classificações corretas, e números fora da diagonal, que representam classificações incorretas.

Percentual corretamente classificado Ver *razão de sucesso*.

Peso discriminante Peso cujo tamanho se relaciona ao poder discriminatório daquela variável independente ao longo dos grupos da variável dependente. Variáveis independentes com grande poder discriminatório geralmente têm pesos grandes, e as que apresentam pouco poder discriminatório geralmente têm pesos pequenos. No entanto, a multicolinearidade entre as variáveis independentes provoca exceções a essa regra. É também chamado de *coeficiente discriminante*.

Pseudo R^2 Um valor de ajuste geral do modelo que pode ser calculado para *regressão logística*; comparável com a medida R^2 usada em regressão múltipla.

Razão de desigualdade A comparação da probabilidade de um evento acontecer com a probabilidade de o evento não acontecer, a qual é usada como uma medida da variável dependente em *regressão logística*.

Razão de sucesso Percentual de objetos (indivíduos, respondentes, empresas etc.) corretamente classificados pela função discriminante. É calculada como o número de objetos na diagonal da *matriz de classificação* dividido pelo número total de objetos. Também conhecida como *percentual corretamente classificado*.

Regressão logística Forma especial de regressão na qual a variável dependente é não-métrica, dicotômica (binária). Apesar de algumas diferenças, a maneira geral de interpretação é semelhante à da regressão linear.

Tolerância Proporção da variação nas variáveis independentes não explicada pelas variáveis que já estão no modelo (função). Pode ser usada como proteção contra a multicolinearidade. Calculada como $1 - R_i^{2*}$, onde R_i^{2*} é a quantia de variância da variável independente *i* explicada por todas as outras variáveis independentes. Uma tolerância de 0 significa que a variável independente sob consideração é uma combinação linear perfeita de variáveis independentes já no modelo. Uma tolerância de 1 significa que uma variável independente é totalmente independente de outras variáveis que já estão no modelo.

Transformação *logit* Transformação dos valores da variável dependente binária discreta da *regressão logística* em uma curva em S (*curva logística*) que representa a probabilidade de um evento. Essa probabilidade é então usada para formar

a *razão de desigualdade,* a qual atua como a variável dependente na regressão logística.

Validação cruzada Procedimento de divisão da amostra em duas partes: a *amostra de análise*, usada na estimação da(s) função(ões) discriminante(s) ou do modelo de *regressão logística*, e a *amostra de teste*, usada para validar os resultados. A validação cruzada evita o super-ajuste da função discriminante ou da regressão logística, permitindo sua validação em uma amostra totalmente separada.

Validação por partição de amostras Ver *validação cruzada.*

Valor de verossimilhança Medida usada em *regressão logística* para representar a falta de ajuste preditivo. Ainda que esses métodos não usem o procedimento dos mínimos quadrados na estimação do modelo, como se faz em regressão múltipla, o valor de verossimilhança é parecido com a soma de erros quadrados na análise de regressão.

Variável categórica Ver *variável não-métrica.*

Variável estatística Combinação linear que representa a soma ponderada de duas ou mais variáveis independentes que formam a *função discriminante.* Também chamada de combinação linear ou composta linear.

Variável métrica Variável com uma unidade constante de medida. Se uma variável métrica tem intervalo de 1 a 9, a diferença entre 1 e 2 é a mesma que aquela entre 8 e 9. Uma discussão mais completa de suas características e diferenças em relação a uma *variável não-métrica* ou *categórica* é encontrada no Capítulo 1.

Variável não-métrica Variável com valores que servem meramente como um rótulo ou meio de identificação, também conhecida como variável categórica, nominal, binária, qualitativa ou taxonômica. O número de um uniforme de futebol é um exemplo. Uma discussão mais completa sobre suas características e diferenças em relação a uma *variável métrica* é encontrada no Capítulo 1.

Vetor Representação da direção e magnitude do papel de uma variável como retratada em uma interpretação gráfica de resultados da análise discriminante.

O QUE SÃO ANÁLISE DISCRIMINANTE E REGRESSÃO LOGÍSTICA?

Ao tentarmos escolher uma técnica analítica apropriada, às vezes encontramos um problema que envolve uma variável dependente categórica e várias variáveis independentes métricas. Por exemplo, podemos querer distinguir riscos de crédito bons de ruins. Se tivéssemos uma medida métrica de risco de crédito, poderíamos usar a regressão múltipla. Em muitos casos não temos a medida métrica necessária para regressão múltipla. Ao invés disso, somos capazes somente de verificar se alguém está em um grupo particular (p.ex., risco de crédito bom ou ruim).

Análise discriminante e regressão logística são as técnicas estatísticas apropriadas quando a variável dependente é **categórica** (nominal ou **não-métrica**) e as **variáveis** independentes são **métricas**. Em muitos casos, a variável dependente consiste em dois grupos ou classificações, por exemplo, masculino versus feminino ou alto versus baixo. Em outros casos, mais de dois grupos são envolvidos, como as classificações em baixo, médio e alto. A análise discriminante é capaz de lidar com dois ou múltiplos (três ou mais) grupos. Quando duas classificações estão envolvidas, a técnica é chamada de *análise discriminante de dois grupos*. Quando três ou mais classificações são identificadas, a técnica é chamada de *análise discriminante múltipla* (*MDA*). A **regressão logística**, também conhecida como **análise *logit***, é limitada, em sua forma básica, a dois grupos, apesar de formulações alternativas poderem lidar com mais de dois grupos.

Análise discriminante

A análise discriminante envolve determinar uma **variável estatística**. Uma variável estatística discriminante é a combinação linear das duas (ou mais) variáveis independentes que melhor discriminarão entre os objetos (pessoas, empresas etc.) nos grupos definidos *a priori.* A discriminação é conseguida estabelecendo-se os pesos da variável estatística para cada variável independente para maximizar as diferenças entre os grupos (i.e., a variância entre grupos relativa à variância interna no grupo). A variável estatística para uma análise discriminante, também conhecida como a **função discriminante**, é determinada a partir de uma equação que se parece bastante com aquela vista em regressão múltipla. Ela assume a seguinte forma:

$$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + \cdots + W_nX_{nk}$$

onde

Z_{jk} = escore Z discriminante da função discriminante j para o objeto k
a = intercepto
W_i = peso discriminante para a variável independente i
X_{ik} = variável independente i para o objeto k

Como acontece com a variável estatística em regressão ou qualquer outra técnica multivariada, percebemos o escore discriminante para cada objeto na análise (pessoa, firma etc.) como sendo uma soma dos valores obtidos pela multiplicação de cada variável independente por seu peso discriminante. O que torna a análise discriminante única é que mais de uma função discriminante pode estar presente, resultando na possibilidade de que cada objeto possa ter mais de um escore discriminante. Discutiremos o que determina o número de funções discriminantes depois, mas aqui vemos que a análise discriminante tem semelhanças e diferenças quando comparada com outras técnicas multivariadas.

A análise discriminante é a técnica estatística apropriada para testar a hipótese de que as médias de grupo de um conjunto de variáveis independentes para dois ou mais grupos são iguais. Calculando a média dos escores

discriminantes para todos os indivíduos em um grupo particular, conseguimos a média do grupo. Essa média de grupo é chamada de **centróide**. Quando a análise envolve dois grupos, há dois centróides; com três grupos, há três centróides, e assim por diante. Os centróides indicam o local mais típico de qualquer indivíduo de um grupo particular, e uma comparação dos centróides de grupos mostra o quão afastados estão os grupos em termos da função discriminante.

O teste para a significância estatística da função discriminante é uma medida generalizada da distância entre os centróides de grupos. Ela é computada comparando-se as distribuições dos escores discriminantes para os grupos. Se a sobreposição nas distribuições é pequena, a função discriminante separa bem os grupos. Se a sobreposição é grande, a função é um discriminador pobre entre os grupos. Duas distribuições de escores discriminantes mostradas na Figura 5-1 ilustram melhor esse conceito. O diagrama do alto representa as distribuições de escores discriminantes para uma função que separa bem os grupos, mostrando sobreposição mínima (a área sombreada) entre os grupos. O diagrama abaixo exibe as distribuições de escores discriminantes em uma função discriminante que é relativamente pobre entre os grupos A e B. As áreas sombreadas de sobreposição representam os casos nos quais podem ocorrer classificação ruim de objetos do grupo A no grupo B e vice-versa.

A análise discriminante múltipla é única em uma característica entre as relações de dependência. Se a variável dependente consiste de mais do que dois grupos, a análise discriminante calcula mais de uma função discriminante. Na verdade, calcula $NG - 1$ funções, onde NG é o número de grupos. Cada função discriminante calcula um escore discriminante Z. No caso de uma variável dependente de três grupos, cada objeto (respondente, empresa etc.) terá um escore separado para funções discriminantes um e dois, permitindo que os objetos sejam representados graficamente em duas dimensões, com cada dimensão representando uma função discriminante. Logo, a análise discriminante não está limitada a uma única variável estatística, como ocorre na regressão múltipla, mas cria múltiplas variáveis estatísticas que representam dimensões de discriminação entre os grupos.

Regressão logística

A regressão logística é uma forma especializada de regressão que é formulada para prever e explicar uma variável categórica binária (dois grupos), e não uma medida dependente métrica. A forma da variável estatística de regressão logística é semelhante à da variável estatística da regressão múltipla. A variável estatística representa uma relação multivariada com coeficientes como os da regressão indicando o impacto relativo de cada variável preditora.

As diferenças entre regressão logística e análise discriminante ficarão mais claras em nossa discussão posterior, neste capítulo, sobre as características únicas da regressão logística. Mas também existem muitas semelhanças entre os dois métodos. Quando as suposições básicas de ambos são atendidas, eles oferecem resultados preditivos e classificatórios comparáveis e empregam medidas diagnósticas semelhantes. A regressão logística, porém, tem a vantagem de ser menos afetada do que a análise discriminante quando as suposições básicas, particularmente a normalidade das variáveis, não são satisfeitas. Ela também pode acomodar variáveis não-métricas por meio da codificação

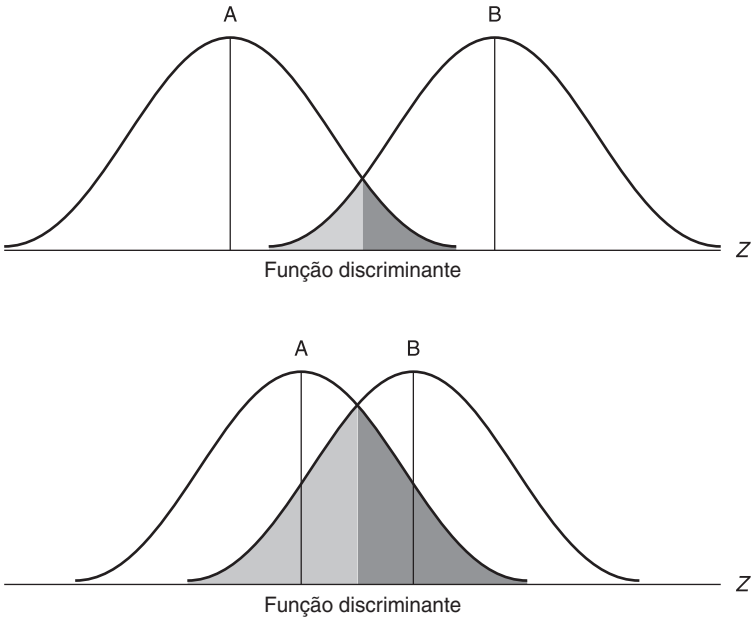

FIGURA 5-1 Representação univariada de escores *Z* discriminantes.

em variáveis dicotômicas, assim como a regressão. No entanto, a regressão logística é limitada a prever apenas uma medida dependente de dois grupos. Logo, em casos nos quais três ou mais grupos formam a medida dependente, a análise discriminante é mais adequada.

ANALOGIA COM REGRESSÃO E MANOVA

A aplicação e interpretação de análise discriminante são quase as mesmas da análise de regressão. Ou seja, a função discriminante é uma combinação linear (variável estatística) de medidas métricas para duas ou mais variáveis independentes e é usada para descrever ou prever uma única variável dependente. A diferença chave é que a análise discriminante é adequada a problemas de pesquisa nos quais a variável dependente é categórica (nominal ou não-métrica), ao passo que a regressão é usada quando a variável dependente é métrica. Como discutido anteriormente, a regressão logística é uma variante da regressão, tendo assim muitas semelhanças, exceto pelo tipo de variável dependente.

A análise discriminante também é comparável à análise multivariada de variância (MANOVA) "reversa", a qual discutimos no Capítulo 6. Na análise discriminante, a variável dependente é categórica e as independentes são métricas. O oposto é verdadeiro em MANOVA, que envolve variáveis dependentes métricas e variável(eis) independente(s) categórica(s). As duas técnicas usam as mesmas medidas estatísticas de ajuste geral do modelo, como será visto a seguir neste e no próximo capítulo.

EXEMPLO HIPOTÉTICO DE ANÁLISE DISCRIMINANTE

A análise discriminante é aplicável a qualquer questão de pesquisa com o objetivo de entender a pertinência a grupos, seja de indivíduos (p. ex., clientes versus não-clientes), empresas (p. ex., lucrativas versus não-lucrativas), produtos (p. ex., de sucesso versus sem sucesso) ou qualquer outro objeto que possa ser avaliado em uma série de variáveis independentes. Para ilustrar as premissas básicas da análise discriminante, examinamos dois cenários de pesquisa, um envolvendo dois grupos (compradores versus não-compradores) e o outro, três grupos (níveis de comportamento de troca). A regressão logística opera de uma maneira comparável à da análise discriminante para dois grupos. Logo, não ilustramos especificamente a regressão logística aqui, adiando nossa discussão até uma consideração separada sobre a regressão logística posteriormente neste capítulo.

Uma análise discriminante de dois grupos: compradores versus não-compradores

Suponha que a KitchenAid queira determinar se um de seus novos produtos – um processador de alimentos novo e aperfeiçoado – será comercialmente bem-sucedido. Ao levar a cabo a investigação, a KitchenAid está interessada em identificar (se possível) os consumidores que comprariam o novo produto e os que não comprariam. Em terminologia estatística, a KitchenAid gostaria de minimizar o número de erros que cometeria ao prever quais consumidores comprariam o novo processador de alimentos e quais não. Para auxiliar na identificação de compradores potenciais, a KitchenAid planejou escalas de avaliação em três características – durabilidade, desempenho e estilo – para serem usadas por consumidores para avaliar o novo produto. Em vez de confiar em cada escala como uma medida separada, a KitchenAid espera que uma combinação ponderada das três preveja melhor se um consumidor tem predisposição para comprar o novo produto.

A meta principal da análise discriminante é obter uma combinação ponderada das três escalas a serem usadas na previsão da possibilidade de um consumidor comprar o produto. Além de determinar se os consumidores que têm tendência para comprar o novo produto podem ser diferenciados daqueles que não têm, a KitchenAid também gostaria de saber quais características de seu novo produto são úteis na diferenciação entre compradores e não-compradores. Ou seja, avaliações de quais das três características do novo produto melhor separam compradores de não-compradores?

Por exemplo, se a resposta "eu compraria" estiver sempre associada com uma medida de alta durabilidade, e a resposta "eu não compraria" estiver sempre associada com uma medida de baixa durabilidade, a KitchenAid concluirá que a característica de durabilidade distingue compradores de não-compradores. Em contrapartida, se a KitchenAid descobrisse que tantas pessoas com alta avaliação para estilo dissessem que comprariam o processador quanto aquelas que não comprariam, então estilo seria uma característica que discrimina muito mal entre compradores e não-compradores.

Identificação de variáveis discriminantes

Para identificar variáveis que possam ser úteis na discriminação entre grupos (ou seja, compradores versus não-compradores), coloca-se ênfase em diferenças de grupos em vez de medidas de correlação usadas em regressão múltipla.

A Tabela 5-1 lista as avaliações dessas três características do novo processador (com um preço especificado) por um painel de 10 compradores em potencial. Ao avaliar o processador de alimentos, cada membro do painel estaria implicitamente comparando-o com produtos já disponíveis no mercado. Depois que o produto foi avaliado, os avaliadores foram solicitados a estabelecer suas intenções de compra ("compraria" ou "não compraria"). Cinco disseram que comprariam o novo processador de alimentos, e cinco disseram que não comprariam.

A Tabela 5-1 identifica diversas variáveis potencialmente discriminantes. Primeiro, uma diferença substancial separa as avaliações médias de X_1 (durabilidade) para os grupos "compraria" e "não compraria" (7,4 versus 3,2). Como tal, a durabilidade parece discriminar bem entre os grupos e ser uma importante característica para compradores em potencial. No entanto, a característica de estilo (X_3) tem uma diferença menor, de 0,2, entre avaliações médias (4,0 – 3,8 = 0,2) para os grupos "compraria" e "não compraria". Portanto, esperaríamos que essa característica fosse menos discriminante em termos de uma decisão de compra. Contudo, antes que possamos fazer tais declarações de forma conclusiva, devemos examinar a distribuição de escores para cada grupo. Desvios-padrão grandes dentro de um ou dos dois grupos podem fazer a diferença entre médias não-significantes e inconseqüente na discriminação entre os grupos.

Como temos apenas 10 respondentes em dois grupos e três variáveis independentes, também podemos olhar os dados graficamente para determinar o que a análise discriminante está tentando conseguir. A Figura 5-2 mostra os dez respondentes em cada uma das três variáveis. O grupo "compraria" é representado por círculos e o grupo "não compraria", por quadrados. Os números de identificação dos respondentes estão dentro das formas.

- X_1 (Durabilidade) tem uma diferença substancial em escores médios, permitindo uma discriminação quase perfeita entre os grupos usando apenas essa variável. Se estabelecêssemos o valor de 5,5 como nosso ponto de corte para discriminar entre os dois grupos, então classificaríamos incorretamente apenas o respondente 5, um dos membros do grupo "compraria". Esta variável ilustra o poder discriminatório ao se ter uma grande diferença nas médias para os dois grupos e uma falta de superposição entre as distribuições dos dois grupos.
- X_2 (Desempenho) fornece uma distinção menos clara entre os dois grupos. No entanto, essa variável dá elevada discriminação para o respondente 5, o qual seria classificado incorretamente se usássemos apenas X_1. Além disso, os respondentes que seriam mal classificados usando X_2 estão bem separados em X_1. Logo, X_1 e X_2 podem efetivamente ser usadas em combinação para prever a pertinência a grupo.
- X_3 (Estilo) mostra pouca distinção entre os grupos. Assim, formando-se uma variável estatística com apenas X_1 e X_2 e omitindo-se X_3, pode-se formar uma função discriminante que maximize a separação dos grupos no escore discriminante.

TABELA 5-1 Resultados do levantamento da KitchenAid para avaliação de um novo produto

	Avaliação do novo produto*		
Grupos baseados em intenção de compra	***X*₁ *Durabilidade***	***X*₂ *Desempenho***	***X*₃ *Estilo***
Grupo 1: Compraria			
Indivíduo 1	8	9	6
Indivíduo 2	6	7	5
Indivíduo 3	10	6	3
Indivíduo 4	9	4	4
Indivíduo 5	4	8	2
Média do grupo	7,4	6,8	4,0
Grupo 2: Não compraria			
Indivíduo 6	5	4	7
Indivíduo 7	3	7	2
Indivíduo 8	4	5	5
Indivíduo 9	2	4	3
Indivíduo 10	2	2	2
Média do grupo	3,2	4,4	3,8
Diferença entre médias de grupos	4,2	2,4	0,2

*Avaliações são feitas em uma escala de 10 pontos (de 1 = muito pobre a 10 = excelente).

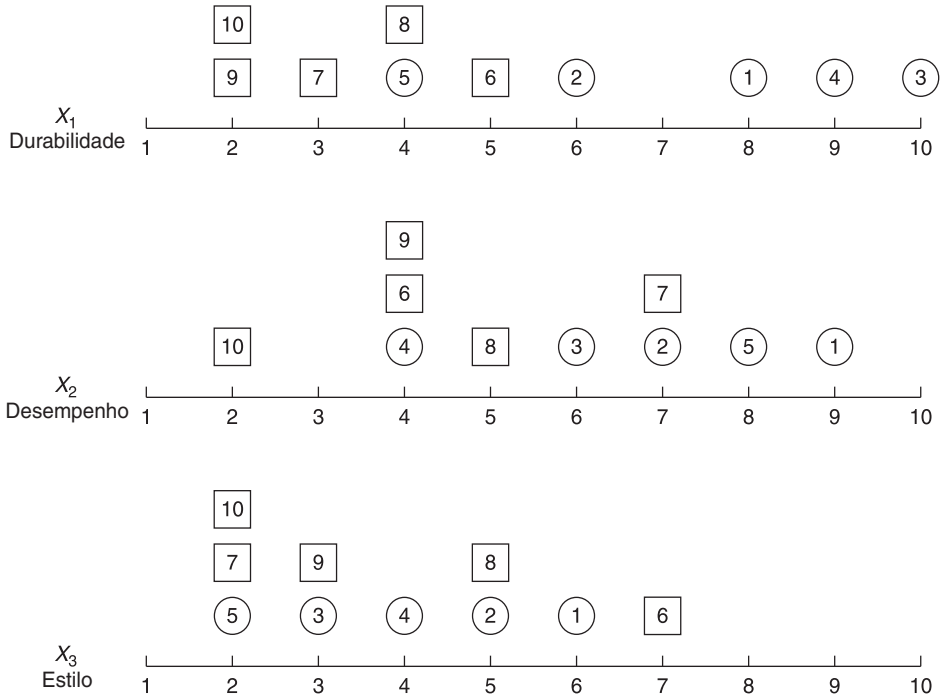

FIGURA 5-2 Representação gráfica de 10 compradores potenciais sobre três variáveis discriminantes possíveis.

Cálculo de uma função discriminante

Com as três variáveis discriminantes potenciais identificadas, a atenção se desvia para a investigação da possibilidade de se usar as variáveis discriminantes em combinação para melhorar o poder discriminatório de qualquer variável individual. Para este fim, uma variável estatística pode ser formada com duas ou mais variáveis discriminantes para atuarem juntas na discriminação entre grupos.

A Tabela 5-2 contém os resultados para três diferentes formulações de funções discriminantes, cada uma representando diferentes combinações das três variáveis independentes.

- A primeira função discriminante contém apenas X_1, igualando o valor de X_1 ao escore discriminante Z (também implicando um peso de 1,0 para X_1 e pesos nulos para as demais variáveis). Como discutido anteriormente, o uso de apenas X_1, o melhor discriminador, resulta na classificação errônea do indivíduo 5, conforme se mostra na Tabela 5-2, onde quatro entre cinco indivíduos do grupo 1 (todos exceto o 5) e cinco entre cinco indivíduos do grupo 2 estão corretamente classificados (i.e., estão na diagonal da matriz de classificação). O percentual corretamente classificado é, portanto, 90% (9 entre 10 sujeitos).
- Como X_2 fornece discriminação para o sujeito 5, podemos formar uma segunda função discriminante combinando igualmente X_1 e X_2 (ou seja, implicando pesos de 1,0 para X_1 e X_2, e 0,0 para X_3) para utilizar os poderes discriminatórios únicos de cada variável. Usando-se um escore de corte de 11 com essa nova função discriminante (ver Tabela 5-2), atinge-se uma perfeita classificação dos dois grupos (100% corretamente classificados). Logo, X_1 e X_2 em combinação são capazes de fazer melhores previsões de pertinência a grupos do que qualquer variável separadamente.
- A terceira função discriminante na Tabela 5-2 representa a verdadeira função discriminante estimada ($Z = -4{,}53 + 0{,}476X_1 + 0{,}359X_2$). Usando um escore de corte de 0, essa terceira função também atinge uma taxa de classificações corretas de 100%, com a máxima separação possível entre os grupos.

Como visto neste exemplo simples, a análise discriminante identifica as variáveis com as maiores diferenças entre os grupos e deriva um coeficiente discriminante que pondera cada variável para refletir tais diferenças. O resultado é uma função discriminante que melhor distingue entre os grupos com base em uma combinação das variáveis independentes.

Uma representação geométrica da função discriminante de dois grupos

Uma ilustração gráfica de uma outra análise de dois grupos ajudará a explicar melhor a natureza da análise discriminante [7]. A Figura 5-3 demonstra o que acontece quando uma função discriminante de dois grupos é computada. Suponha que temos dois grupos, A e B, e duas medidas, V_1 e V_2, para cada membro dos dois grupos. Podemos representar graficamente em um diagrama de dispersão a associação da variável V_1 com a variável V_2 para cada membro dos dois grupos. Na Figura 5-3, os

TABELA 5-2 Criação de funções discriminantes para prever compradores versus não-compradores

	Escores Z discriminantes calculados		
Grupo	**Função 1** $Z = X_1$	**Função 2** $Z = X_1 + X_2$	**Função 3** $Z = -4{,}53 + 0{,}476X_1 + 0{,}359X_2$
Grupo 1: Compraria			
Indivíduo 1	8	17	2,51
Indivíduo 2	6	13	0,84
Indivíduo 3	10	16	2,38
Indivíduo 4	9	13	1,19
Indivíduo 5	4	12	0,25
Grupo 2: Não compraria			
Indivíduo 6	5	9	–0,71
Indivíduo 7	3	10	–0,59
Indivíduo 8	4	9	–0,83
Indivíduo 9	2	6	–2,14
Indivíduo 10	2	4	–2,86
Escore de corte	5,5	11	0,0

Precisão de classificação

	Grupo previsto		Grupo previsto		Grupo previsto	
Grupo real	1	2	1	2	1	2
1: Compraria	4	1	5	0	5	0
2: Não-compraria	0	5	0	5	0	5

pontos pequenos* representam as medidas das variáveis para os membros do grupo B, e os pontos grandes* correspondem ao grupo A. As elipses desenhadas em torno dos pontos pequenos e grandes envolveriam alguma proporção pré-especificada dos pontos, geralmente 95% ou mais em cada grupo. Se desenharmos uma reta pelos dois pontos nos quais as elipses se interceptam e então projetarmos a reta sobre um novo eixo Z, podemos dizer que a sobreposição entre as distribuições univariadas A′ e B′ (representada pela área sombreada) é menor do que se fosse obtida por qualquer outra reta através das elipses formadas pelos diagramas de dispersão [7].

O importante a ser notado a respeito da Figura 5-3 é que o eixo Z expressa os perfis de duas variáveis dos grupos A e B como números únicos (escores discriminantes). Encontrando uma combinação linear das variáveis originais V_1 e V_2, podemos projetar os resultados como uma função discriminante. Por exemplo, se os pontos pequenos e grandes são projetados sobre o novo eixo Z como escores Z discriminantes, o resultado condensa a informação sobre diferenças de grupos (mostrada no gráfico V_1V_2) em um conjunto de pontos (escores Z) sobre um único eixo, mostrado pelas distribuições A′ e B′.

Para resumir, para um dado problema de análise discriminante, uma combinação linear das variáveis independentes é determinada, resultando em uma série de escores discriminantes para cada objeto em cada grupo. Os escores discriminantes são computados de acordo com a regra estatística de maximizar a variância entre os grupos e minimizar a variância dentro deles. Se a variância entre os grupos é grande em relação à variância dentro dos grupos, dizemos que a função discriminante separa bem os grupos.

* N. de R. T.: Na verdade, os pontos nos grupos A e B não diferem em tamanho e, sim, no formato. No A a forma é quadrada e no B é circular.

Um exemplo de análise discriminante de três grupos: intenções de troca

O exemplo de dois grupos já examinado demonstra o objetivo e o benefício de se combinarem variáveis independentes em uma variável estatística para fins de discriminação entre grupos. A análise discriminante também tem um outro meio de discriminação – a estimação e o uso de múltiplas variáveis estatísticas – em casos onde há três ou mais grupos. Essas funções discriminantes agora se tornam dimensões de discriminação, sendo cada dimensão separada e diferente da outra. Assim, além de melhorar a explicação de pertinência ao grupo, essas funções discriminantes adicionais dão informação quanto às várias combinações de variáveis independentes que discriminam entre grupos.

Para ilustrar uma aplicação de análise discriminante a três grupos, examinamos a pesquisa conduzida pela HBAT referente à possibilidade de os clientes de um concorrente trocarem de fornecedores. Um pré-teste em pequena escala envolveu entrevistas de 15 clientes de um concorrente importante. Durante as entrevistas, os clientes foram indagados sobre a probabilidade de trocarem

(*Continua*)

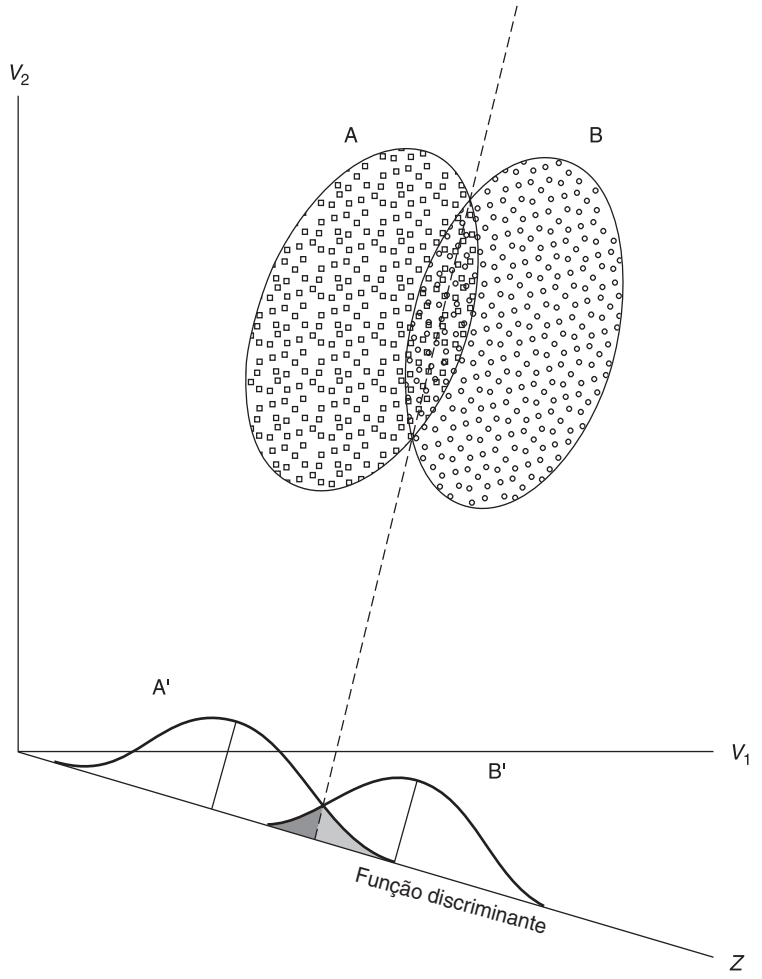

FIGURA 5-3 Ilustração gráfica da análise discriminante de dois grupos.

(*Continuação*)

de fornecedores em uma escala de três categorias. As três respostas possíveis eram "definitivamente trocaria", "indeciso" e "definitivamente não trocaria". Clientes foram designados a grupos 1, 2 ou 3, respectivamente, de acordo com suas respostas. Os clientes também avaliaram o concorrente em duas características: competitividade de preço (X_1) e nível de serviço (X_2). A questão da pesquisa agora é determinar se as avaliações dos clientes a respeito do concorrente podem prever sua probabilidade de trocar de fornecedor. Como a variável dependente de troca de fornecedor foi medida como uma variável categórica (não-métrica) e as medidas de preço e serviço são métricas, a análise discriminante é adequada.

Identificação de variáveis discriminantes

Com três categorias da variável dependente, a análise discriminante pode estimar duas funções discriminantes, cada uma representando uma dimensão diferente de discriminação.

A Tabela 5-3 contém os resultados da pesquisa para os 15 clientes, cinco em cada categoria da variável dependente. Como fizemos no exemplo de dois grupos, podemos olhar para os escores médios de cada grupo para ver se uma das variáveis discrimina bem entre todos os grupos. Para X_1, competitividade de preço, percebemos uma grande diferença de médias entre o grupo 1 e os grupos 2 ou 3 (2,0 versus 4,6 ou 3,8). X_1 pode discriminar bem entre o grupo 1 e os grupos 2 ou 3, mas é muito menos eficiente para discriminar entre os grupos 2 e 3. Para X_2, nível de serviço, percebemos que a diferença entre os grupos 1 e 2 é muito pequena (2,0 versus 2,2), ao passo que há uma grande diferença entre o grupo 3 e os grupos 1 ou 2 (6,2 versus 2,0 ou 2,2). Logo, X_1 distingue o grupo 1 dos grupos 2 e 3, e X_2 diferencia o grupo 3 dos grupos 1 e 2. Como resultado, vemos que X_1 e X_2 fornecem diferentes "dimensões" de discriminação entre os grupos.

TABELA 5-3 Resultados da pesquisa HBAT sobre intenções de troca por clientes potenciais

	*Avaliação do fornecedor atual**	
Grupos baseados em intenção de troca	X_1 **Competitividade de preço**	X_2 **Nível do serviço**
Grupo 1: Definitivamente trocaria		
Indivíduo 1	2	2
Indivíduo 2	1	2
Indivíduo 3	3	2
Indivíduo 4	2	1
Indivíduo 5	2	3
Média do grupo	2,0	2,0
Grupo 2: Indeciso		
Indivíduo 6	4	2
Indivíduo 7	4	3
Indivíduo 8	5	1
Indivíduo 9	5	2
Indivíduo 10	5	3
Média do grupo	4,6	2,2
Grupo 3: Definitivamente não trocaria		
Indivíduo 11	2	6
Indivíduo 12	3	6
Indivíduo 13	4	6
Indivíduo 14	5	6
Indivíduo 15	5	7
Média do grupo	3,8	6,2

*Avaliações são feitas em uma escala de 10 pontos (de 1 = muito pobre a 10 = excelente).

Cálculo de duas funções discriminantes

Com as potenciais variáveis discriminantes identificadas, o próximo passo é combiná-las em funções discriminantes que utilizarão seu poder combinado de diferenciação para separar grupos.

Para ilustrar graficamente essas dimensões, a Figura 5-4 retrata os três grupos em cada variável independente separadamente. Vendo os membros dos grupos em qualquer variável, podemos perceber que nenhuma variável discrimina bem entre todos os grupos. Mas se construímos duas funções discriminantes simples, usando apenas pesos simples de 1,0 e 0,0, os resultados se tornam muito mais claros. A função discriminante 1 dá para X_1 um peso de 1,0, e para X_2 um peso de 0,0. Do mesmo modo, a função discriminante 2 dá para X_2 um peso de 1,0 e para X_1 um peso de 0,0. As funções podem ser enunciadas matematicamente como

$$\text{Função discriminante } 1 = 1{,}0(X_1) + 0{,}0(X_2)$$
$$\text{Função discriminante } 2 = 0{,}0(X_1) + 1{,}0(X_2)$$

Essas equações mostram em termos simples como o procedimento de análise discriminante estima os pesos para maximizar a discriminação.

Com as duas funções, agora podemos calcular dois escores discriminantes para cada respondente. Além disso, as duas funções discriminantes fornecem as dimensões de discriminação.

A Figura 5-4 também contém um gráfico de cada respondente em uma representação bidimensional. A separação entre grupos agora fica bastante clara, e cada grupo pode ser facilmente diferenciado. Podemos estabelecer valores em cada dimensão que definirão regiões contendo cada grupo (p.ex., todos os membros do grupo 1 estão na região menos que 3,5 na dimensão 1 e menos que 4,5 na dimensão 2). Cada um dos outros grupos pode ser analogamente definido em termos das amplitudes dos escores de suas funções discriminantes.

Em termos de dimensões de discriminação, a primeira função discriminante, competitividade de preço, diferencia clientes indecisos (mostrados com um quadrado) de clientes que decidiram trocar (círculos). Mas competitividade de preço não diferencia aqueles que decidiram não trocar (losangos). Em vez disso, a percepção de nível de serviço, que define a segunda função discriminante, prevê se um cliente decidirá não trocar versus se um cliente está indeciso ou determinado a trocar de forne-

(*Continua*)

(*Continuação*)
cedores. O pesquisador pode apresentar à gerência os impactos separados de competitividade de preço e nível de serviço para a tomada de decisões.

A estimação de mais de uma função discriminante, quando possível, fornece ao pesquisador uma discriminação melhorada e perspectivas adicionais sobre as características e as combinações que melhor discriminam entre os grupos. As seções a seguir detalham os passos necessários

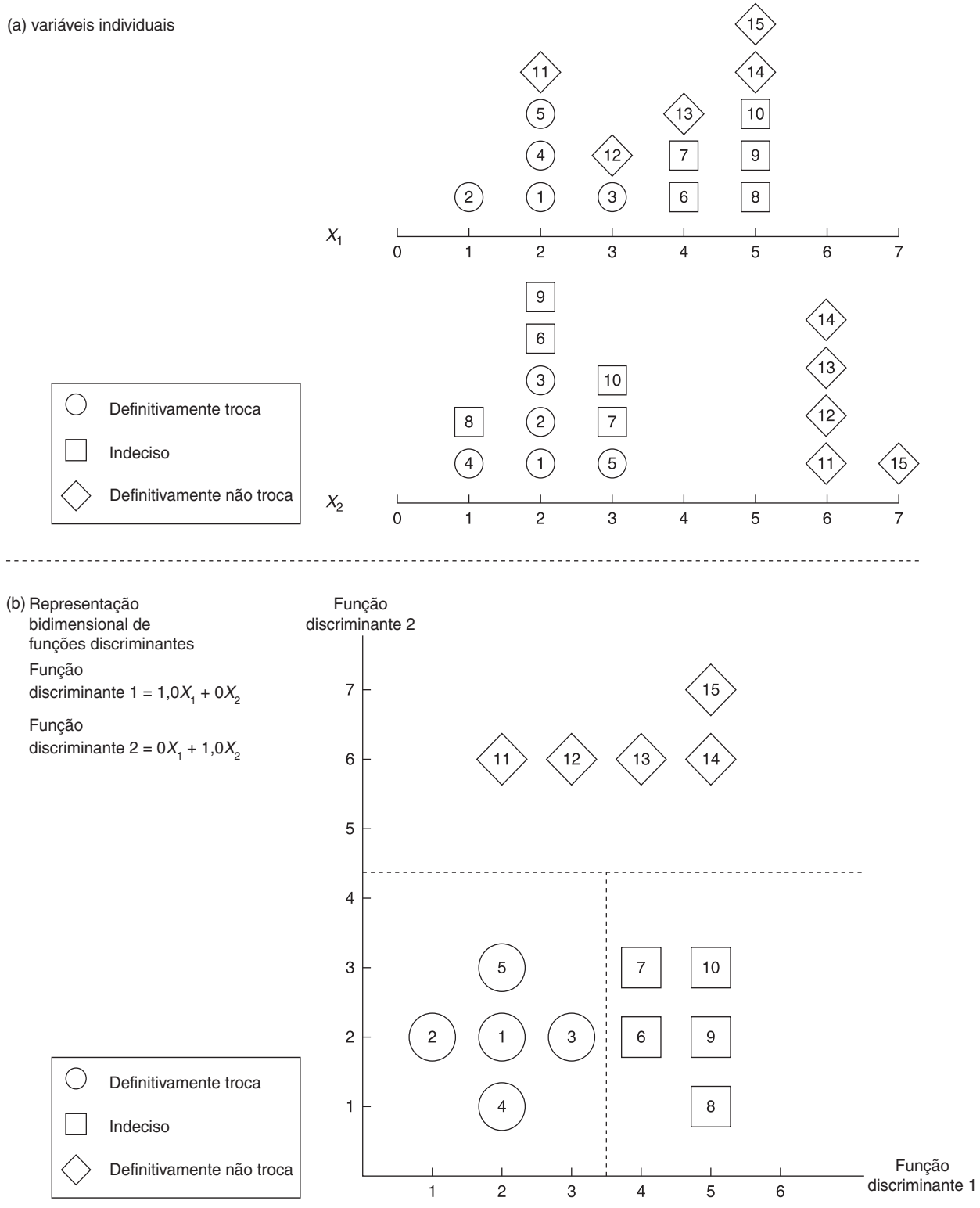

FIGURA 5-4 Representação gráfica de variáveis discriminantes potenciais para uma análise discriminante de três grupos.

para se executar uma análise discriminante, avaliar seu nível de ajuste preditivo e então interpretar a influência de variáveis independentes ao se fazer uma previsão.

O PROCESSO DE DECISÃO PARA ANÁLISE DISCRIMINANTE

A aplicação de análise discriminante pode ser vista da perspectiva da construção de modelo de seis estágios introduzida no Capítulo 1 e retratada na Figura 5-5 (estágios 1-3) e na Figura 5-6 (estágios 4-6). Assim como em todas as aplicações multivariadas, estabelecer os objetivos é o primeiro passo na análise. Em seguida, o pesquisador deve abordar questões específicas de planejamento e se certificar de que as suposições inerentes estão sendo atendidas. A análise continua com a dedução da função discriminante e a determinação de se uma função estatisticamente significante pode ser obtida para separar os dois (ou mais) grupos. Os resultados discriminantes são então avaliados quanto à precisão preditiva pelo desenvolvimento de uma matriz de classificação. Em seguida, a interpretação da função discriminante determina qual das variáveis independentes mais contribui para discriminar entre os grupos. Finalmente, a função discriminante deve ser validada com uma amostra de teste. Cada um desses estágios é discutido nas seções a seguir. Discutimos a regressão logística em uma seção à parte depois de examinarmos o processo de decisão para a análise discriminante. Desse modo, as semelhanças e diferenças entre essas duas técnicas podem ser destacadas.

ESTÁGIO 1: OBJETIVOS DA ANÁLISE DISCRIMINANTE

Uma revisão dos objetivos de aplicar a análise discriminante deve esclarecer melhor sua natureza. A análise discriminante pode abordar qualquer um dos seguintes objetivos de pesquisa:

1. Determinar se existem diferenças estatisticamente significantes entre os perfis de escore médio em um conjunto de variáveis para dois (ou mais) grupos definidos *a priori*.
2. Determinar quais das variáveis independentes explicam o máximo de diferenças nos perfis de escore médio dos dois ou mais grupos.
3. Estabelecer o número e a composição das dimensões de discriminação entre grupos formados a partir do conjunto de variáveis independentes.
4. Estabelecer procedimentos para classificar objetos (indivíduos, firmas, produtos e assim por diante) em grupos, com base em seus escores em um conjunto de variáveis independentes.

Como observado nesses objetivos, a análise discriminante é útil quando o pesquisador está interessado em compreender diferenças de grupos ou em classificar obje-

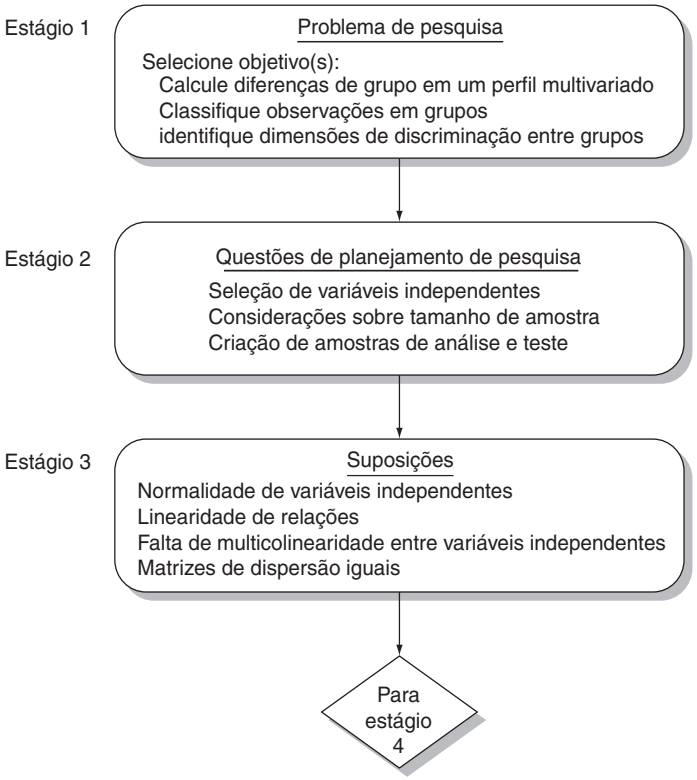

FIGURA 5-5 Estágios 1-3 no diagrama de decisão da análise discriminante.

tos corretamente em grupos ou classes. Portanto, a análise discriminante pode ser considerada um tipo de análise de perfil ou uma técnica preditiva analítica. Em qualquer caso, a técnica é mais apropriada onde existe uma só variável dependente categórica e diversas variáveis independentes métricas.

- Como uma *análise de perfil*, a análise discriminante fornece uma avaliação objetiva de diferenças entre grupos em um conjunto de variáveis independentes. Nesta situação, a análise discriminante é bastante semelhante à análise multivariada de variância (ver Capítulo 6 para uma discussão mais detalhada de análise multivariada de variância). Para entender as diferenças de grupos, a análise discriminante permite discernir o papel de variáveis individuais, bem como definir combinações dessas variáveis que representam dimensões de discriminação entre grupos. Essas dimensões são os efeitos coletivos de diversas variáveis que trabalham conjuntamente para distinguir entre os grupos. O uso de métodos de estimação seqüenciais também permite identificar subconjuntos de variáveis com o maior poder discriminatório.
- Para *fins de classificação*, a análise discriminante fornece uma base para classificar não somente a amostra usada para estimar a função discriminante, mas também quaisquer outras observações que possam ter valores para todas as variáveis independentes. Desse modo, a análise discriminante pode ser usada para classificar outras observações nos grupos definidos.

ESTÁGIO 2: PROJETO DE PESQUISA PARA ANÁLISE DISCRIMINANTE

A aplicação bem-sucedida da análise discriminante requer a consideração de várias questões. Tais questões incluem a seleção da variável dependente e das variáveis independentes, o tamanho necessário da amostra para a estimação das funções discriminantes, e a divisão da amostra para fins de validação.

Seleção de variáveis dependente e independentes

Para aplicar a análise discriminante, o pesquisador deve primeiramente especificar quais variáveis devem ser independentes e qual deve ser a medida dependente. Lembre-se que a variável dependente é categórica e as independentes são métricas.

A variável dependente

O pesquisador deve se concentrar na variável dependente primeiro. O número de grupos (categorias) da variável dependente pode ser dois ou mais, mas esses grupos devem ser mutuamente excludentes e cobrir todos os casos. Ou seja, cada observação pode ser colocada em apenas um grupo. Em alguns casos, a variável dependente pode envolver dois grupos (dicotômicas), como bom versus ruim. Em outros casos, a variável dependente envolve vários grupos (multicotômica), como as ocupações de médico, advogado ou professor.

Quantas categorias na variável dependente? Teoricamente, a análise discriminante pode lidar com um número ilimitado de categorias na variável dependente. Na prática, porém, o pesquisador deve selecionar uma variável dependente e o número de categorias com base em diversas considerações.

1. Além de serem mutuamente excludentes e exaustivas, as categorias da variável dependente devem ser distintas e únicas no conjunto escolhido de variáveis independentes. A análise discriminante considera que cada grupo *deveria* ter um perfil único nas variáveis independentes usadas, e assim desenvolve as funções discriminantes para separar ao máximo os grupos com base nessas variáveis. Não obstante, a análise discriminante não tem um meio para acomodar ou combinar categorias que não sejam distintas nas variáveis independentes. Se dois ou mais grupos têm perfis semelhantes, a análise discriminante não será capaz de estabelecer univocamente o perfil de cada grupo, resultando em uma explicação e classificação mais pobres dos grupos como um todo. Dessa forma, o pesquisador deve escolher as variáveis dependentes e suas categorias para refletir diferenças nas variáveis independentes. Um exemplo ajudará a ilustrar este ponto.

> Imagine que o pesquisador deseja identificar diferenças entre categorias ocupacionais baseado em algumas características demográficas (p.ex., renda, formação, características familiares). Se ocupações fossem representadas por um pequeno número de categorias (p.ex., pessoal de segurança e limpeza, técnicos, pessoal de escritório e profissionais de nível superior), então esperaríamos que houvesse diferenças únicas entre os grupos e que a análise discriminante seria mais adequada para desenvolver funções discriminantes que explicariam as distinções de grupos e classificariam com sucesso os indivíduos em suas categorias corretas.
>
> Se, porém, o número de categorias ocupacionais fosse aumentado, a análise discriminante poderia ter uma dificuldade maior para identificar diferenças. Por exemplo, considere que a categoria de profissionais de nível superior fosse expandida para as categorias de médicos, advogados, gerentes gerais, professores universitários e assim por diante. A despeito de esta expansão fornecer uma classificação ocupacional mais refinada, seria muito mais difícil fazer distinções entre essas categorias com base em variáveis demográficas. Os resultados teriam um desempenho mais pobre na análise discriminante, tanto em termos de explicação quanto de classificação.

2. O pesquisador deve também buscar um número menor, e não maior, de categorias na medida dependente. Pode parecer mais lógico expandir o número de categorias em busca de mais agrupamentos únicos, mas a expansão do número

de categorias apresenta mais complexidades nas tarefas de classificação e estabelecimento de perfil na análise discriminante. Se a análise discriminante pode estimar $NG - 1$ (número de grupos menos um) funções discriminantes, então o aumento do número de grupos expande o número de possíveis funções discriminantes, aumentando a complexidade da identificação das dimensões inerentes de discriminação refletidas por conta de cada função discriminante, bem como representando o efeito geral de cada variável independente.

Como esses dois pontos sugerem, o pesquisador sempre deve equilibrar a vontade de expandir as categorias em favor da unicidade (exclusividade) com a crescente efetividade de um número menor de categorias. O pesquisador deve testar e selecionar uma variável dependente com categorias que tenham as maiores diferenças entre todos os grupos, ao mesmo tempo que mantenham suporte conceitual e relevância administrativa.

Conversão de variáveis métricas Os exemplos anteriores de variáveis categóricas eram verdadeiras dicotomias (ou multicotomias). Há algumas situações, contudo, em que a análise discriminante é apropriada mesmo se a variável dependente não é verdadeiramente categórica (não-métrica). Podemos ter uma variável dependente de medida ordinal ou intervalar, a qual queremos usar como uma variável dependente categórica. Em tais casos, teríamos de criar uma variável categórica, e duas abordagens estão entre as mais usuais:

- O método mais comum é estabelecer categorias usando uma escala métrica. Por exemplo, se tivéssemos uma variável que medisse o número médio de refrigerantes consumidos por dia e os indivíduos respondessem em uma escala de zero a oito ou mais por dia, poderíamos criar uma tricotomia (três grupos) artificial simplesmente designando aqueles indivíduos que consumissem nenhum, um ou dois refrigerantes por dia como usuários modestos, aqueles que consumissem três, quatro ou cinco por dia como usuários médios, e os que consumissem seis, sete, oito ou mais como usuários pesados. Tal procedimento criaria uma variável categórica de três grupos na qual o objetivo seria discriminar entre usuários de refrigerantes que fossem modestos, médios e pesados. Qualquer número de grupos categóricos artificiais pode ser desenvolvido. Mais freqüentemente, a abordagem envolveria a criação de duas, três ou quatro categorias. Um número maior de categorias poderia ser estabelecido se houvesse necessidade.
- Quando três ou mais categorias são criadas, surge a possibilidade de se examinarem apenas os grupos extremos em uma análise discriminante de dois grupos. A **abordagem de extremos polares** envolve a comparação somente dos dois grupos extremos e a exclusão do grupo do meio da análise discriminante. Por exemplo, o pesquisador poderia examinar os usuários modestos e pesados de refrigerantes e excluir os usuários médios. Esse tratamento pode ser usado toda vez que o pesquisador desejar olhar apenas os grupos extremos. Contudo, ele também pode querer tentar essa abordagem quando os resultados de uma análise de regressão não são tão bons quanto o previsto. Tal procedimento pode ser útil porque é possível que diferenças de grupos possam aparecer até quando os resultados de regressão são pobres. Ou seja, a abordagem de extremos polares com a análise discriminante pode revelar diferenças que não são tão evidentes em uma análise de regressão do conjunto completo de dados [7]. Tal manipulação dos dados naturalmente necessitaria de cuidado na interpretação das descobertas.

As variáveis independentes

Depois de ter tomado uma decisão sobre a variável dependente, o pesquisador deve decidir quais variáveis independentes serão incluídas na análise. As variáveis independentes geralmente são selecionadas de duas maneiras. A primeira abordagem envolve a identificação de variáveis a partir de pesquisa prévia ou do modelo teórico que é a base inerente da questão de pesquisa. A segunda abordagem é a intuição – utilizar o conhecimento do pesquisador e selecionar intuitivamente variáveis para as quais não existe pesquisa prévia ou teoria, mas que logicamente poderiam ser relacionadas à previsão dos grupos para a variável dependente.

Em ambos os casos, as variáveis independentes mais apropriadas são aquelas que diferem da variável dependente em pelo menos dois dos grupos. Lembre que o propósito de qualquer variável independente é apresentar um perfil único de pelo menos um grupo quando comparado a outros. Variáveis que não diferem ao longo dos grupos são de pouca utilidade em análise discriminante.

Tamanho da amostra

A análise discriminante, como as outras técnicas multivariadas, é afetada pelo tamanho da amostra sob análise. Como discutido no Capítulo 1, amostras muito pequenas têm grandes erros amostrais, de modo que a identificação de todas, exceto as grandes diferenças, é improvável. Além disso, amostras muito grandes tornarão todas as diferenças estatisticamente significantes, ainda que essas mesmas diferenças possam ter pouca ou nenhuma relevância administrativa. Entre esses extremos, o pesquisador deve considerar o impacto do tamanho das amostras sobre a análise discriminante, tanto no nível geral quanto em uma base de grupo-por-grupo.

Tamanho geral da amostra

A primeira consideração envolve o tamanho geral da amostra. A análise discriminante é bastante sensível à proporção do tamanho da amostra em relação ao número de variáveis preditoras. Como resultado, muitos estudos sugerem uma proporção de 20 observações para cada variável preditora. Apesar de essa proporção poder ser difícil de manter na prática, o pesquisador deve notar que os resultados se tornam instáveis quando o tamanho da amostra diminui em relação ao número de variáveis independentes. O tamanho mínimo recomendado é de cinco

observações por variável independente. Note que essa proporção se aplica a todas as variáveis consideradas na análise, mesmo que todas as variáveis consideradas não entrem na função discriminante (como na estimação *stepwise*).

Tamanho da amostra por categoria

Além do tamanho da amostra geral, o pesquisador também deve considerar o tamanho da amostra de cada categoria. No mínimo, o menor grupo de uma categoria deve exceder o número de variáveis independentes. Como uma orientação prática, cada categoria deve ter no mínimo 20 observações. Mas mesmo que todas as categorias excedam 20 observações, o pesquisador também deve considerar os tamanhos relativos das mesmas. Se os grupos variam muito em tamanho, isso pode causar impacto na estimação da função discriminante e na classificação de observações. No estágio de classificação, grupos maiores têm uma chance desproporcionalmente maior de classificação. Se os tamanhos de grupos variam muito, o pesquisador pode querer extrair uma amostra aleatoriamente a partir do(s) grupo(s) maior(es), reduzindo assim seu(s) tamanho(s) a um nível comparável ao(s) grupo(s) menor(es). Sempre se lembre, porém, de manter um tamanho adequado de amostra geral e para cada grupo.

Divisão da amostra

Uma observação final sobre o impacto do tamanho da amostra na análise discriminante. Como será posteriormente discutido no estágio 6, a maneira preferida de validar uma análise discriminante é dividir a amostra em duas sub-amostras, uma usada para estimação da função discriminante e outra para fins de validação. Em termos de considerações sobre tamanho amostral, é essencial que cada sub-amostra tenha tamanho adequado para suportar as conclusões dos resultados. Dessa forma, todas as considerações discutidas na seção anterior se aplicam não somente à amostra total, mas agora a cada uma das duas sub-amostras (especialmente aquela usada para estimação). Nenhuma regra rígida e rápida foi desenvolvida, mas parece lógico que o pesquisador queira pelo menos 100 na amostra total para justificar a divisão da mesma em dois grupos.

Criação das sub-amostras

Vários procedimentos têm sido sugeridos para dividir a amostra em sub-amostras. O procedimento usual é dividir a amostra total de respondentes aleatoriamente em dois grupos. Um deles, a **amostra de análise**, é usado para desenvolver a função discriminante. O segundo grupo, a **amostra de teste**, é usado para testar a função discriminante. Esse método de validação da função é chamado de abordagem de **partição da amostra** ou **validação cruzada** [1,5,9,18].

Nenhuma orientação definitiva foi estabelecida para determinar os tamanhos relativos das sub-amostras de análise e de teste (ou validação). O procedimento mais popular é dividir a amostra total de forma que metade dos respondentes seja colocada na amostra de análise e a outra metade na amostra de teste. No entanto, nenhuma regra rígida e rápida foi estabelecida, e alguns pesquisadores preferem uma partição 60-40 ou mesmo 75-25 entre os grupos de análise e de teste, dependendo do tamanho da amostra geral.

Quando se selecionam as amostras de análise e teste, geralmente segue-se um procedimento de amostragem proporcionalmente estratificada. Assuma primeiro que o pesquisador deseja uma divisão 50-50. Se os grupos categóricos para a análise discriminante são igualmente representados na amostra total, as amostras de estimação e de teste devem ser de tamanhos aproximadamente iguais. Se os grupos originais são diferentes, os tamanhos das amostras de estimação e de teste devem ser proporcionais em relação à distribuição da amostra total. Por exemplo, se uma amostra consiste em 50 homens e 50 mulheres, as amostras de estimação e de teste teriam 25 homens e 25 mulheres cada. Se a amostra tiver 70 mulheres e 30 homens, então as amostras de estimação e de teste consistirão em 35 mulheres e 15 homens cada.

E se a amostra geral for muito pequena?

Se a amostra é muito pequena para justificar uma divisão em grupos de análise e de teste, o pesquisador tem duas opções. Primeiro, desenvolver a função na amostra inteira e então usar a função para classificar o mesmo grupo usado para desenvolver a função. Esse procedimento resulta em um viés ascendente na precisão preditiva da função, mas certamente é melhor do que não testar a função de forma alguma. Segundo, diversas técnicas discutidas no estágio 6 podem desempenhar um tipo de procedimento de teste no qual a função discriminante é repetidamente estimada sobre a amostra, cada vez reservando uma observação diferente para previsão. Nesta abordagem, amostras muito menores podem ser usadas, pois a amostra geral não precisa ser dividida em sub-amostras.

ESTÁGIO 3: SUPOSIÇÕES DA ANÁLISE DISCRIMINANTE

Como ocorre em todas as técnicas multivariadas, a análise discriminante é baseada em uma série de suposições. Tais suposições se relacionam a processos estatísticos envolvidos nos procedimentos de estimação e classificação e a questões que afetam a interpretação dos resultados. A seção a seguir discute cada tipo de suposição e os impactos sobre a aplicação apropriada da análise discriminante.

Impactos sobre estimação e classificação

As suposições-chave para determinar a função discriminante são a de normalidade multivariada das variáveis independentes, e a de estruturas (matrizes) de dispersão e covariância desconhecidas (mas iguais) para os grupos como definidos pela variável dependente [8,10]. Existem evidências da sensibilidade da análise discriminante a violações dessas suposições. Os testes para normalidade discutidos no Capítulo 2 estão disponíveis ao pesquisador, juntamente com o teste **M de Box** para avaliar a similaridade das matrizes de dispersão das variáveis independentes entre os grupos. Se as suposições são violadas, o pesquisador deve considerar métodos alternativos (p.ex., regressão logística, descrita na próxima seção) e compreender os impactos sobre os resultados que podem ser esperados.

Impacto sobre estimação

Dados que não atendem a suposição de normalidade multivariada podem causar problemas na estimação da função discriminante. Ações corretivas podem ser viáveis através de transformações dos dados para reduzir as disparidades entre as matrizes de covariância. No entanto, em muitos casos essas ações corretivas são ineficientes. Em tais casos, os modelos devem ser diretamente validados. Se a medida dependente é binária, a regressão logística deve ser utilizada sempre que possível.

Impacto sobre classificação

Matrizes de covariância desiguais também afetam negativamente o processo de classificação. Se os tamanhos das amostras são pequenos e as matrizes de covariância são diferentes, então a significância estatística do processo de estimação é afetada adversamente. O caso mais comum é o de covariâncias desiguais entre grupos de tamanho adequado, em que as observações são super-classificadas nos grupos com matrizes de covariância maiores. Esse efeito pode ser minimizado aumentando-se o tamanho da amostra e também usando-se as matrizes de covariância específicas dos grupos para fins de classificação, mas essa abordagem exige a validação cruzada dos resultados discriminantes. Finalmente, técnicas de classificação quadráticas estão disponíveis em muitos dos programas estatísticos caso existam grandes diferenças entre as matrizes de covariância dos grupos e as ações corretivas não minimizem o efeito [6,12,14].

Impactos sobre interpretação

Uma outra característica dos dados que afeta os resultados é a multicolinearidade entre as variáveis independentes. A multicolinearidade, medida em termos de **tolerância**, denota que duas ou mais variáveis independentes estão altamente correlacionadas, de modo que uma variável pode ser altamente explicada ou prevista pela(s) outra(s) variável(eis), acrescentando pouco ao poder explicativo do conjunto como um todo. Essa consideração se torna especialmente crítica quando procedimentos *stepwise* são empregados. O pesquisador, ao interpretar a função discriminante, deve estar ciente da multicolinearidade e de seu impacto na determinação de quais variáveis entram na solução *stepwise*. Para uma discussão mais detalhada da multicolinearidade e seu impacto nas soluções *stepwise*, ver o Capítulo 4. Os procedimentos para detectar a presença da multicolinearidade são também abordados no Capítulo 4.

Como em qualquer técnica multivariada que emprega uma variável estatística, uma suposição implícita é a de que todas as relações são lineares. As relações não-lineares não são refletidas na função discriminante, a menos que transformações específicas de variáveis sejam executadas para representarem efeitos não-lineares. Finalmente, observações atípicas podem ter um impacto substancial na precisão de classificação de quaisquer resultados da aná-

REGRAS PRÁTICAS 5-1

Planejamento de análise discriminante

- A variável dependente deve ser não-métrica, representando grupos de objetos que devem diferir nas variáveis independentes
- Escolha uma variável dependente que:
  - Melhor represente diferenças de grupos de interesse
  - Defina grupos que são substancialmente distintos
  - Minimize o número de categorias ao mesmo tempo que atenda aos objetivos da pesquisa
- Ao converter variáveis métricas para uma escala não-métrica para uso como a variável dependente, considere o uso de grupos extremos para maximizar as diferenças de grupos
- Variáveis independentes devem identificar diferenças entre pelo menos dois grupos para uso em análise discriminante
- A amostra deve ser grande o bastante para:
  - Ter pelo menos uma observação a mais por grupo do que o número de variáveis independentes, mas procurar por pelo menos 20 casos por grupo
  - Ter 20 casos por variável independente, com um nível mínimo recomendado de 5 observações por variável
  - Ter uma amostra grande o bastante para dividi-la em amostras de teste e de estimação, cada uma atendendo às exigências acima
- A suposição mais importante é a igualdade das matrizes de covariância, o que afeta tanto estimação quanto classificação
- Multicolinearidade entre as variáveis independentes pode reduzir sensivelmente o impacto estimado de variáveis independentes na função discriminante derivada, particularmente no caso de emprego de um processo de estimação *stepwise*

lise discriminante. O pesquisador é encorajado a examinar todos os resultados quanto à presença de observações atípicas e a eliminar observações atípicas verdadeiras, se necessário. Para uma discussão sobre algumas das técnicas que avaliam as violações das suposições estatísticas básicas ou a detecção de observações atípicas, ver Capítulo 2.

ESTÁGIO 4: ESTIMAÇÃO DO MODELO DISCRIMINANTE E AVALIAÇÃO DO AJUSTE GERAL

Para determinar a função discriminante, o pesquisador deve decidir o método de estimação e então determinar o número de funções a serem retidas (ver Figura 5-6). Com as funções estimadas, o ajuste geral do modelo pode ser avaliado de diversas maneiras. Primeiro, **escores *Z* discriminantes**, também conhecidos como os **escores *Z***, podem ser calculados para cada objeto. A comparação das médias dos grupos (centróides) nos escores *Z* fornece uma medida de discriminação entre grupos. A precisão preditiva pode ser medida como o número de observações classificadas nos grupos corretos, com vários critérios disponíveis para avaliar se o processo de classificação alcança significância prática ou estatística. Finalmente, diagnósticos por casos podem identificar a precisão de classificação de cada caso e seu impacto relativo sobre a estimação geral do modelo.

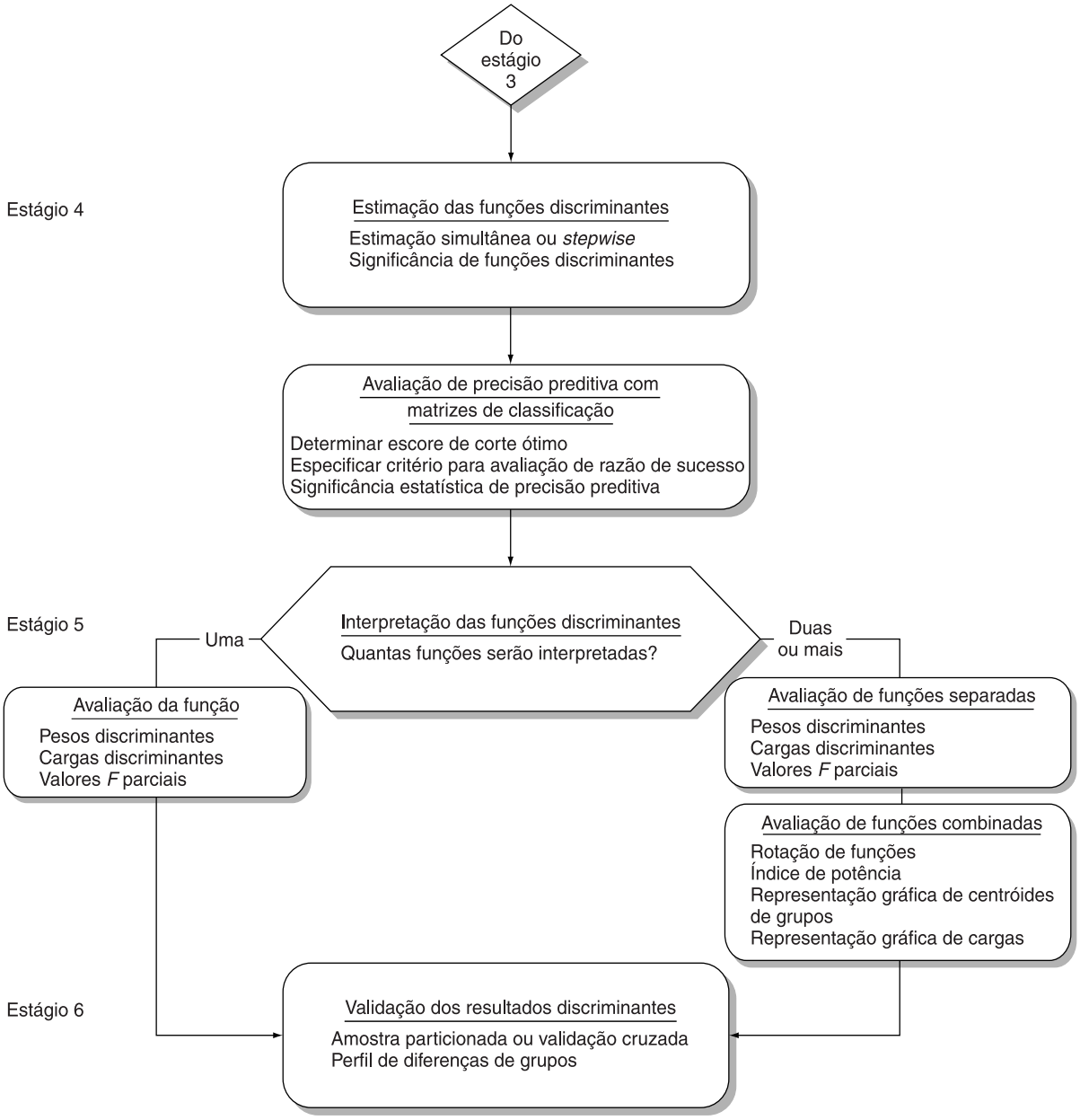

FIGURA 5-6 Estágios 4-6 no diagrama de decisão da análise discriminante.

Seleção de um método de estimação

A primeira tarefa na obtenção da função discriminante é selecionar o método de estimação. Ao fazer tal escolha, o pesquisador deve balancear a necessidade de controle sobre o processo de estimação com o desejo pela parcimônia nas funções discriminantes. Os dois métodos disponíveis são o simultâneo (direto) e o *stepwise*, cada um discutido adiante.

Estimação simultânea

A **estimação simultânea** envolve a computação da função discriminante, de modo que todas as variáveis independentes são consideradas juntas. Assim, a função discriminante é computada com base no conjunto inteiro de variáveis independentes, sem consideração do poder discriminatório de cada uma delas. O método simultâneo é apropriado quando, por conta de razões teóricas, o pesquisador quer incluir todas as variáveis independentes na análise e não está interessado em ver resultados intermediários baseados apenas nas variáveis mais discriminantes.

Estimação stepwise

A **estimação *stepwise*** é uma alternativa à abordagem simultânea. Envolve a inclusão das variáveis independentes na função discriminante, uma por vez, com base em seu poder discriminatório. A abordagem *stepwise* segue um processo seqüencial de adicionar ou descartar variáveis da seguinte maneira:

1. Escolher a melhor variável discriminatória.
2. Comparar a variável inicial com cada uma das outras variáveis independentes, uma de cada vez, e selecionar a variável mais adequada para melhorar o poder discriminatório da função em combinação com a primeira variável.
3. Selecionar as demais variáveis de maneira semelhante. Note que conforme variáveis adicionais são incluídas, algumas previamente escolhidas podem ser removidas se a informação que elas contêm sobre diferenças de grupos estiver disponível em alguma combinação das outras variáveis incluídas em estágios posteriores.
4. Considerar o processo concluído quando todas as variáveis independentes forem incluídas na função ou as variáveis excluídas forem julgadas como não contribuindo significantemente para uma discriminação futura.

O método *stepwise* é útil quando o pesquisador quer considerar um número relativamente grande de variáveis independentes para inclusão na função. Selecionando-se seqüencialmente a próxima melhor variável discriminante em cada passo, as variáveis que não são úteis na discriminação entre os grupos são eliminadas e um conjunto reduzido de variáveis é identificado. O conjunto reduzido geralmente é quase tão bom quanto – e às vezes melhor que – o conjunto completo de variáveis.

O pesquisador deve notar que a estimação *stepwise* se torna menos estável e generalizável à medida que a proporção entre tamanho da amostra e variável independente diminui abaixo do nível recomendado de 20 observações por variável independente. É particularmente importante, nesses casos, validar os resultados de tantas maneiras quanto possível.

Significância estatística

Após a estimação da função discriminante, o pesquisador deve avaliar o nível de significância para o poder discriminatório coletivo das funções discriminantes, bem como a significância de cada função discriminante em separado. A avaliação da significância geral fornece ao pesquisador a informação necessária para decidir se deve proceder com a interpretação da análise ou se uma reespecificação se faz necessária. Se o modelo geral for significante, a avaliação das funções individuais identifica aquelas que devem ser mantidas e interpretadas.

Significância geral

Ao se avaliar a significância estatística do modelo geral, diferentes critérios são aplicáveis para procedimentos de estimação simultânea versus *stepwise*. Em ambas as situações, os testes estatísticos se relacionam com a habilidade das funções discriminantes de obterem escores Z discriminantes que sejam significantemente diferentes entre grupos.

Estimação simultânea. Quando uma abordagem simultânea é usada, as medidas de lambda de Wilks, o traço de Hotelling e o critério de Pillai avaliam a significância estatística do poder discriminatório da(s) função(ões) discriminante(s). A maior raiz característica de Roy avalia apenas a primeira função discriminante. Para uma discussão mais detalhada sobre as vantagens e desvantagens de cada critério, veja a discussão de testes de significância em análise multivariada de variância no Capítulo 6.

Estimação *stepwise*. Se um método *stepwise* é empregado para estimar a função discriminante, as medidas D^2 de Mahalanobis e V de Rao são mais adequadas. Ambas são medidas de distância generalizada. O procedimento D^2 de Mahalanobis é baseado em distância euclideana quadrada generalizada que se adapta a variâncias desiguais. A maior vantagem deste procedimento é que ele é computado no espaço original das variáveis preditoras, em vez de ser computado como uma versão extraída de outras medidas. O procedimento D^2 de Mahalanobis se torna particularmente crítico quando o número de variáveis preditoras aumenta porque ele não resulta em redução de dimensionalidade. Uma perda em dimensionalidade causaria uma perda de informação, porque ela diminui a variabilidade das variáveis independentes. Em geral, D^2 de Mahalanobis é o procedimento preferido quando o pesquisador está interessado no uso máximo de informação disponível em um processo *stepwise*.

Significância de funções discriminantes individuais

Se o número de grupos é três ou mais, então o pesquisador deve decidir não apenas se a discriminação entre grupos é estatisticamente significante, mas também se cada função discriminante estimada é estatisticamente significante. Como discutido anteriormente, a análise discriminante estima uma função discriminante a menos do que o número de grupos. Se três grupos são analisados, então duas funções discriminantes serão estimadas; para quatro grupos, três funções serão estimadas, e assim por diante. Todos os programas de computador fornecem ao pesquisador a informação necessária para verificar o número de funções necessárias para obter significância estatística, sem incluir funções discriminantes que não aumentam o poder discriminatório significantemente.

O critério de significância convencional de 0,05 ou acima é freqüentemente usado, sendo que alguns pesquisadores estendem o nível requerido (p.ex., 0,10 ou mais) com base na ponderação de custo versus o valor da informação. Se os maiores níveis de risco para incluir resultados não-significantes (p.ex., níveis de significância > 0,05) são aceitáveis, pode-se reter funções discriminantes que são significantes no nível 0,2 ou até mesmo 0,3.

Se uma ou mais funções são consideradas estatisticamente não-significantes, o modelo discriminante deve ser reestimado com o número de funções a serem determinadas limitado ao número de funções significantes. Desse modo, a avaliação de precisão preditiva e a interpretação das funções discriminantes serão baseadas apenas em funções significantes.

Avaliação do ajuste geral do modelo

Logo que as funções discriminantes significantes tenham sido identificadas, a atenção se desvia para a verificação do ajuste geral das funções discriminantes mantidas. Essa avaliação envolve três tarefas:

1. Calcular escores *Z* discriminantes para cada observação
2. Calcular diferenças de grupos nos escores *Z* discriminantes
3. Avaliar a precisão de previsão de pertinência a grupos.

Devemos observar que o emprego da função discriminante para fins de classificação é apenas um entre dois possíveis tratamentos. O segundo utiliza uma **função de classificação**, também conhecida como **função discriminante linear de Fisher**. As funções de classificação, uma para cada grupo, são usadas exclusivamente para classificar observações. Nesse método de classificação, os valores de uma observação para as variáveis independentes são inseridos nas funções de classificação, e um escore de classificação para cada grupo é calculado para aquela observação. A observação é então classificada no grupo com o maior escore de classificação.

Examinamos a função discriminante como o meio de classificação porque ela fornece uma representação concisa e simples de cada função discriminante, simplificando o processo de interpretação e a avaliação da contribuição de variáveis independentes. Ambos os métodos conseguem resultados comparáveis, apesar de usarem diferentes meios.

Cálculo de escores Z discriminantes

Com as funções discriminantes retidas definidas, a base para calcular os escores *Z* discriminantes foi estabelecida. Como discutido anteriormente, o escore *Z* discriminante de qualquer função discriminante pode ser calculado para cada observação pela seguinte fórmula:

$$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + \cdots + W_nX_{nk}$$

onde

Z_{jk} = escore Z discriminante da função discriminante j para o objeto k
a = intercepto
W_i = coeficiente discriminante para a variável independente i
X_{ik} = variável independente i para o objeto k

Este escore, uma variável métrica, fornece uma maneira direta de comparar observações em cada função. Assume-se que as observações com escores *Z* semelhantes são mais parecidas com base nas variáveis que constituem essa função do que aquelas com escores totalmente distintos. A função discriminante pode ser expressa com pesos e valores padronizados ou não-padronizados. A versão padronizada é mais útil para fins de interpretação, mas a não-padronizada é mais fácil de utilizar no cálculo do escore *Z* discriminante.

Avaliação de diferenças de grupos

Uma vez que os escores *Z* discriminantes são calculados, a primeira avaliação de ajuste geral do modelo é determinar a magnitude de diferenças entre os membros de cada grupo em termos dos escores *Z* discriminantes. Uma

REGRAS PRÁTICAS 5-2

Estimação e ajuste do modelo

- Apesar de a estimação *stepwise* poder parecer ótima ao selecionar o mais parcimonioso conjunto de variáveis maximamente discriminantes, cuidado com o impacto de multicolinearidade sobre a avaliação do poder discriminatório de cada variável.
- O ajuste geral do modelo avalia a significância estatística entre grupos sobre os escores *Z* discriminantes, mas não avalia precisão preditiva.
- Tendo mais de dois grupos, não confine sua análise a apenas as funções discriminantes estatisticamente significantes, mas considere a possibilidade de funções não-significantes (com níveis de até 0,3) adicionarem poder explanatório.

medida resumo das diferenças de grupos é uma comparação dos **centróides** dos grupos, o escore *Z* discriminante médio para todos os membros dos grupos. Uma medida de sucesso da análise discriminante é sua habilidade em definir função(ões) discriminante(s) que resulte(m) em centróides de grupos significantemente diferentes. As diferenças entre centróides são medidas em termos do D^2 de Mahalanobis, para o qual há testes disponíveis para determinar se as diferenças são estatisticamente significantes. O pesquisador deve garantir que, mesmo com funções discriminantes significantes, há diferenças consideráveis entre os grupos.

Os centróides de grupos em cada função discriminante também podem ser representados graficamente para demonstrar os resultados de uma perspectiva gráfica. Gráficos geralmente são preparados para as primeiras duas ou três funções discriminantes (assumindo que elas são funções estatisticamente significantes). Os valores para cada grupo mostram sua posição no espaço discriminante reduzido (assim chamado porque nem todas as funções e, assim, nem toda a variância, são representadas graficamente). O pesquisador pode ver as diferenças entre os grupos em cada função; no entanto, a inspeção visual não explica totalmente o que são essas diferenças. Pode-se desenhar círculos que envolvam a distribuição de observações em volta de seus respectivos centróides para esclarecer melhor as diferenças de grupos, mas esse procedimento está além do escopo deste texto (ver Dillon e Goldstein [4]).

Avaliação da precisão preditiva de pertinência a grupo

Dado que a variável dependente é não-métrica, não é possível usar uma medida como R^2, como se faz em regressão múltipla, para avaliar a precisão preditiva. Em vez disso, cada observação deve ser avaliada com o objetivo de saber se ela foi corretamente classificada. Ao fazer isso, diversas considerações importantes devem ser feitas:

- A concepção estatística e prática para desenvolver matrizes de classificação
- A determinação do escore de corte
- A construção das matrizes de classificação
- Os padrões para avaliar a precisão de classificação

Por que matrizes de classificação são desenvolvidas. Os testes estatísticos para avaliar a significância das funções discriminantes somente avaliam o grau de diferença entre os grupos com base nos escores *Z* discriminantes, mas não dizem quão bem a função prevê. Esses testes estatísticos sofrem das mesmas desvantagens dos testes de hipóteses clássicos. Por exemplo, suponha que os dois grupos são considerados significantemente diferentes além do nível 0,01. Com amostras suficientemente grandes, as médias de grupo (centróides) poderiam ser virtualmente idênticas e ainda teriam significância estatística. Para determinar a habilidade preditiva de uma função discriminante, o pesquisador deve construir matrizes de classificação.

A **matriz de classificação** fornece uma perspectiva sobre significância prática, e não sobre significância estatística. Com a análise discriminante múltipla, o percentual **corretamente classificado**, também conhecido como **razão de sucesso**, revela o quão bem a função discriminante classificou os objetos. Com uma amostra suficientemente grande em análise discriminante, poderíamos ter uma diferença estatisticamente significante entre os dois (ou mais) grupos e mesmo assim classificar corretamente apenas 53% (quando a chance é de 50%, com grupos de mesmo tamanho) [16]. Em tais casos, o teste estatístico indicaria significância estatística, ainda que a razão de sucesso viabilizasse um julgamento à parte a ser feito em termos de significância prática. Logo, devemos usar o procedimento da matriz de classificação para avaliar precisão preditiva além de simples significância estatística.

Cálculo do escore de corte. Usando as funções discriminantes consideradas significantes, podemos desenvolver matrizes de classificação para uma avaliação mais precisa do poder discriminatório das funções. Antes que uma matriz de classificação seja definida, porém, o pesquisador deve determinar o **escore de corte** (também chamado de valor *Z* crítico) para cada função discriminante. O escore de corte é o critério em relação ao qual o escore discriminante de cada objeto é comparado para determinar em qual grupo o objeto deve ser classificado.

O escore de corte representa o ponto divisor usado para classificar observações em um entre dois grupos baseado no escore da função discriminante. O cálculo de um escore de corte entre dois grupos quaisquer é baseado nos centróides de dois grupos (média de grupo dos escores discriminantes) e no tamanho relativo dos grupos. Os centróides são facilmente calculados e fornecidos em cada estágio do processo *stepwise*. Para calcular corretamente o **escore de corte ótimo**, o pesquisador deve abordar dois pontos:

1. Definir as probabilidades *a priori*, baseado nos tamanhos relativos dos grupos observados ou especificados pelo pesquisador (ou assumidos iguais, ou com valores dados pelo pesquisador).
2. Calcular o valor do escore de corte ótimo como uma média ponderada sobre os tamanhos assumidos dos grupos (obtido a partir das probabilidades *a priori*).

Definição das probabilidades a priori. O impacto e a importância de tamanhos relativos de grupos são muitas vezes desconsiderados, apesar de serem baseados nas suposições do pesquisador relativas à representatividade da amostra. Neste caso, representatividade se relaciona à representação dos tamanhos relativos dos grupos na população real, o que pode ser estabelecido como probabili-

dades *a priori* (ou seja, a proporção relativa de cada grupo em relação à amostra total).

A questão fundamental é: os tamanhos relativos dos grupos são representativos dos tamanhos de grupos na população? A suposição padrão para a maioria dos programas estatísticos é de probabilidades iguais; em outras palavras, cada grupo é considerado como tendo a mesma chance de ocorrer, mesmo que os tamanhos dos grupos na amostra sejam desiguais. Se o pesquisador está inseguro sobre se as proporções observadas na amostra são representativas das proporções da população, a abordagem conservadora é empregar probabilidades iguais. Em alguns casos, estimativas das probabilidades *a priori* podem estar disponíveis, como em pesquisa anterior. Aqui a suposição padrão de probabilidades iguais *a priori* é substituída por valores especificados pelo pesquisador. Em qualquer caso, os reais tamanhos de grupos são substituídos com base nas probabilidades *a priori* especificadas.

No entanto, se a amostra foi conduzida aleatoriamente e o pesquisador sente que os tamanhos de grupos são representativos da população, então o pesquisador pode especificar probabilidade *a priori* com base na amostra de estimação. Assim, os verdadeiros tamanhos de grupos são considerados representativos e diretamente usados no cálculo do escore de corte (ver a discussão que se segue). Em todos os casos, porém, o pesquisador deve especificar como as probabilidades *a priori* são calculadas, o que afeta os tamanhos de grupos usados no cálculo como ilustrado.

Por exemplo, considere uma amostra de teste consistindo de 200 observações, com tamanhos de grupos de 60 a 140 que se relacionam com probabilidades *a priori* de 30% e 70%, respectivamente. Se a amostra é considerada representativa, então os tamanhos de 60 e 140 são empregados no cálculo do escore de corte. Não obstante, se a amostra é considerada não-representativa, o pesquisador deve especificar as probabilidades *a priori*. Se elas são especificadas como iguais (50% e 50%), os tamanhos amostrais de 100 e 100 seriam usados no cálculo do escore de corte no lugar dos tamanhos reais. Especificar outros valores para as probabilidades *a priori* resultaria em diferentes tamanhos amostrais para os dois grupos.

Cálculo do escore de corte ótimo. A importância das probabilidades *a priori* no escore de corte é muito evidente depois que se percebe como o mesmo é calculado. A fórmula básica para computar o escore de corte entre dois grupos quaisquer é:

$$Z_{CS} = \frac{N_A Z_B + N_B Z_A}{N_A + N_B}$$

onde

Z_{CS} = escore de corte ótimo entre grupos A e B
N_A = número de observações no grupo A
N_B = número de observações no grupo B
Z_A = centróide para o grupo A
Z_B = centróide para o grupo B

Com tamanhos desiguais de grupos, o escore de corte ótimo para uma função discriminante é agora a média ponderada dos centróides de grupos. O escore de corte é ponderado na direção do grupo menor, gerando, com sorte, uma melhor classificação do grupo maior.

Se os grupos são especificados como sendo de iguais tamanhos (probabilidades *a priori* definidas como iguais), então o escore de corte ótimo estará a meio caminho entre os dois centróides e se torna simplesmente a média dos mesmos:

$$Z_{CE} = \frac{Z_A + Z_B}{2}$$

onde

Z_{CE} = valor do escore de corte crítico para grupos de mesmo tamanho
Z_A = centróide do grupo A
Z_B = centróide do grupo B

Ambas as fórmulas para cálculo do escore de corte ótimo assumem que as distribuições são normais e as estruturas de dispersão de grupos são conhecidas.

O conceito de um escore de corte ótimo para grupos iguais e distintos é ilustrado nas Figuras 5-7 e 5-8, respectivamente. Os escores de corte ponderados e não-ponderados são mostrados. Fica evidente que se o grupo A é muito menor que o grupo B, o escore de corte ótimo está mais próximo ao centróide do grupo A do que ao centróide do grupo B. Além disso, se o escore de corte não-ponderado fosse usado, nenhum dos objetos no grupo A seria mal classificado, mas uma parte substancial dos que estão no grupo B seria mal classificada.

Custos de má classificação. O escore de corte ótimo também deve considerar o custo de classificar um objeto no grupo errado. Se os custos de má classificação são aproximadamente iguais para todos os grupos, o escore de corte ótimo será aquele que classificar mal o menor número de objetos em todos os grupos. Se os custos de má classificação são desiguais, o escore de corte ótimo será o que minimizar os custos de má classificação. Abordagens mais sofisticadas para determinar escores de corte são discutidas em Dillon e Goldstein [4] e Huberty et al. [13]. Essas abordagens são baseadas em um modelo estatístico bayesiano e são adequadas quando os custos de má classi-

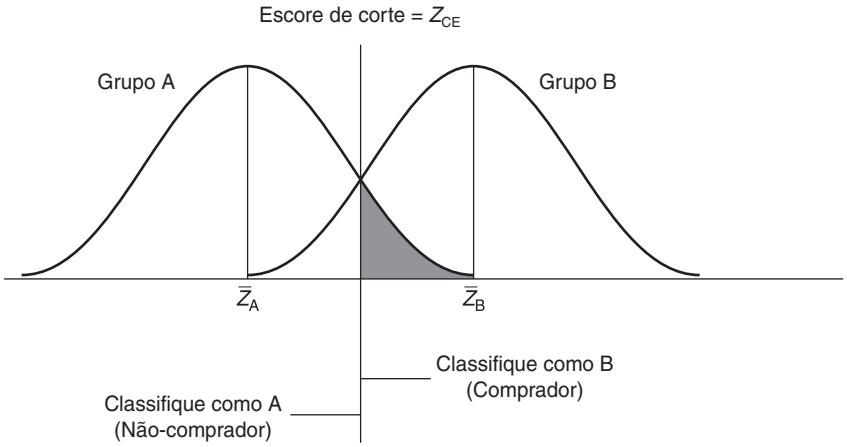

FIGURA 5-7 Escore de corte ótimo com amostras de tamanhos iguais.

ficação em certos grupos são altos, quando os grupos são de tamanhos muito diferentes, ou quando se deseja tirar vantagem de um conhecimento *a priori* de probabilidades de pertinência a grupo.

Na prática, quando se calcula o escore de corte, geralmente não é necessário inserir as medidas originais da variável para cada indivíduo na função discriminante e obter o escore discriminante para cada pessoa para usar no cálculo de Z_A e Z_B (centróides dos grupos A e B). O programa de computador fornece os escores discriminantes, bem como Z_A e $Z_{B,}$ como *output* regular. Quando o pesquisador tem os centróides de grupo e os tamanhos da amostra, o escore de corte ótimo pode ser obtido simplesmente substituindo-se os valores na fórmula apropriada.

Construção das matrizes de classificação. Para validar a função discriminante pelo uso de matrizes de classificação, a amostra deve ser aleatoriamente dividida em dois grupos. Um dos grupos (a amostra de análise) é usado para computar a função discriminante. O outro (a amostra de teste ou de validação) é retido para uso no desenvolvimento da matriz de classificação. O procedimento envolve a multiplicação dos pesos gerados pela amostra de análise pelas medidas originais da variável da amostra de teste. Em seguida, os escores discriminantes individuais para a amostra de teste são comparados com o valor do escore de corte crítico e classificados como se segue:

Classifique um indivíduo no grupo A se $Z_n < Z_{ct}$

ou

Classifique um indivíduo no grupo B se $Z_n > Z_{ct}$.

onde

Z_n = escore Z discriminante para o n-ésimo indivíduo
Z_{ct} = valor do escore de corte crítico

Os resultados do procedimento de classificação são apresentados em forma matricial, como mostrado na Tabela 5-4. As entradas na diagonal da matriz representam o número de indivíduos corretamente classificados. Os números fora da diagonal representam as classificações incorretas. As entradas sob a coluna rotulada de "Tamanho do grupo real" representam o número de indivíduos que realmente estão em cada um dos dois grupos. As entradas na base das colunas representam o número de indivíduos designados aos grupos pela função discriminante. O per-

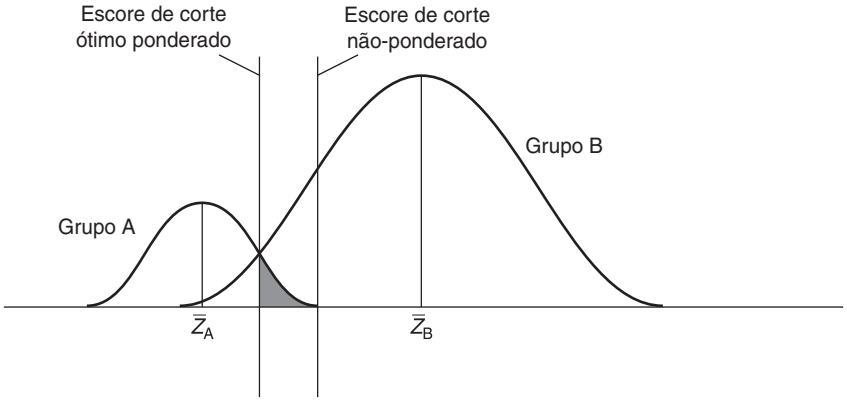

FIGURA 5-8 Escore de corte ótimo com tamanhos desiguais de amostras.

centual corretamente classificado para cada grupo é mostrado no lado direito da matriz, e o percentual geral corretamente classificado, também conhecido como a razão de sucesso, é mostrado na base.

> Em nosso exemplo, o número de indivíduos corretamente designados ao grupo 1 é 22, enquanto 3 membros do grupo 1 estão incorretamente designados ao grupo 2. Do mesmo modo, o número de classificações corretas no grupo 2 é 20, e o número de designações incorretas no grupo 1 é 5. Assim, os percentuais de precisão de classificação da função discriminante para os grupos reais 1 e 2 são 88% e 80%, respectivamente. A precisão de classificação geral (razão de sucesso) é 84%.

Um tópico final sobre os procedimentos de classificação é o teste *t* disponível para determinar o nível de significância para a precisão de classificação. A fórmula para uma análise de dois grupos (igual tamanho de amostra) é

$$t = \frac{p - 0{,}5}{\sqrt{\frac{0{,}5(1{,}0 - 0{,}5)}{N}}}$$

onde

p = proporção corretamente classificada
N = tamanho da amostra

Essa fórmula pode ser adaptada para uso com mais grupos e diferentes tamanhos de amostra.

Estabelecimento de padrões de comparação para a razão de sucesso. Como observado anteriormente, a precisão preditiva da função discriminante é medida pela razão de sucesso, a qual é obtida a partir da matriz de classificação. O pesquisador pode questionar o que é ou não considerado um nível aceitável de precisão preditiva para uma função discriminante. Por exemplo, 60% é um nível aceitável ou deveríamos esperar obter de 80% a 90% de precisão preditiva? Para responder essa questão o pesquisador deve primeiro determinar o percentual que poderia ser classificado corretamente por *chances* (*sem a ajuda da função discriminante*).

Padrões de comparação para a razão de sucesso em grupos de mesmo tamanho. Quando os tamanhos de amostra dos grupos são iguais, a determinação da classificação por chances é bem simples; ela é obtida dividindo-se 1 pelo número de grupos. A fórmula é

$$C_{\text{IGUAL}} = 1/(\text{Número de grupos}).$$

Por exemplo, para uma função de dois grupos, a probabilidade seria de 0,50; para uma função de três grupos, seria de 0,33, e assim por diante.

Padrões de comparação para a razão de sucesso em grupos de tamanhos desiguais. A determinação da classificação por chances para situações nas quais os tamanhos dos grupos são desiguais é um pouco mais complicada. Devemos considerar apenas o maior grupo, a probabilidade combinada de todos os tamanhos de grupos, ou algum outro padrão? Imaginemos que temos uma amostra total de 200 indivíduos divididos como amostras de teste e de análise de 100 observações cada. Na amostra de teste, 75 objetos pertencem a um grupo e 25 ao outro. Examinaremos os possíveis caminhos nos quais podemos construir um padrão para comparação e aquilo que cada um representa.

- Conhecido como o **critério de chance máxima**, poderíamos arbitrariamente designar todos os indivíduos ao maior grupo. O critério da chance máxima deve ser usado quando o único objetivo da análise discriminante é maximizar o percentual corretamente classificado [16]. É também o padrão mais conservador, pois ele gera o mais alto padrão de comparação. No entanto, são raras as situações nas quais estamos interessados apenas em maximizar o percentual corretamente classificado. Geralmente, o pesquisador usa a análise discriminante para identificar corretamente os membros de todos os grupos. Em casos nos quais os tamanhos das amostras são desiguais e o pesquisador deseja classificar os membros de todos os grupos, a função discriminante vai contra as chances, classificando um indivíduo no(s) grupo(s) menor(es). O critério por chances não leva esse fato em consideração [16].

TABELA 5-4 Matriz de classificação para análise discriminante de dois grupos

	Grupo previsto			
Grupo real	**1**	**2**	**Tamanho do grupo real**	**Percentual corretamente classificado**
1	22	3	25	88
2	5	20	25	80
Tamanho previsto do grupo	27	23	50	84[a]

[a]Percentual corretamente classificado = (Número corretamente classificado/Número total de observações) x 100
= [(22 + 20)/50] × 100
= 84%

Em nosso exemplo simples de uma amostra com dois grupos (75 e 25 pessoas cada), usando esse método teríamos uma precisão de classificação de 75%, o que se conseguiria classificando-se todos no grupo maior sem a ajuda de qualquer função discriminante. Pode-se concluir que, a menos que a função discriminante consiga uma precisão de classificação maior do que 75%, ela deve ser descartada, pois não nos ajuda a melhorar a precisão preditiva que podemos atingir sem qualquer análise discriminante.

- Quando os tamanhos de grupos são desiguais e o pesquisador deseja identificar corretamente os membros de todos os grupos, não apenas do maior, o **critério de chances proporcionais** é considerado por muitos como o mais apropriado. A fórmula para esse critério é

$$C_{\text{PRO}} = p^2 + (1 - p)^2$$

onde

p = proporção de indivíduos no grupo 1
$1 - p$ = proporção de indivíduos no grupo 2

Usando os tamanhos de grupos de nosso exemplo anterior (75 e 25), percebemos que o critério de chances proporcionais seria de 62,5% $[0{,}75^2 + (1{,}0 - 0{,}75)^2 = 0{,}625]$ comparado com 75%. Logo, neste caso, uma precisão preditiva de 75% seria aceitável porque está acima dos 62,5% do critério de chances proporcionais.

- Um problema dos critérios de chance máxima e de chances proporcionais são os tamanhos das amostras usados para cálculo dos padrões. Você deve usar grupos com o tamanho da amostra geral, da amostra de análise/estimação, ou da amostra de validação/teste? Aqui vão algumas sugestões:
  - Se os tamanhos das amostras de análise e estimação são considerados suficientemente grandes (i.e., amostra total de 100 com cada grupo tendo pelo menos 20 casos), obtenha padrões separados para cada amostra.
  - Se as amostras separadas não são consideradas suficientemente grandes, use os tamanhos de grupos da amostra total para calcular os padrões.
  - Atente a tamanhos de grupos diferentes entre amostras quando usar o critério de chance máxima, pois ele depende do maior tamanho de grupo. Esta orientação é especialmente crítica quando a amostra é pequena ou quando as proporções de tamanhos de grupos variam muito de amostra para amostra. Este é outro motivo de cautela no emprego do critério de chance máxima.

Esses critérios de chances são úteis somente quando computados com amostras de teste (abordagem da partição da amostra). Se os indivíduos usados no cálculo da função discriminante são os classificados, o resultado é um viés ascendente na precisão preditiva. Em tais casos, os critérios deveriam ser ajustados para cima em função desse viés.

Comparação da razão de sucesso com o padrão. A questão de "quanta precisão de classificação devo ter?" é crucial. Se o percentual de classificações corretas é significantemente maior do que se esperaria por chances, o pesquisador pode proceder à interpretação das funções discriminantes e de perfis de grupos. No entanto, se a precisão de classificação não é maior do que pode ser esperado das chances, quaisquer diferenças que pareçam existir merecem pouca ou nenhuma interpretação; ou seja, as diferenças em perfis de escores não forneceriam qualquer informação significativa para identificar a pertinência a grupos.

A questão, então, é o quanto a precisão de classificação deve ser relativa às chances? Por exemplo, se as chances são de 50% (dois grupos, com iguais tamanhos), uma precisão de classificação (preditiva) de 60% justifica ir para o estágio de interpretação? Em última instância, a decisão depende do custo em relação ao valor da informação. O argumento do custo versus valor oferece pouca ajuda ao pesquisador iniciante, mas o seguinte critério é sugerido: *A precisão de classificação deve ser pelo menos um quarto maior do que a obtida por chances.*

Por exemplo, se a precisão por chances for de 50%, a precisão de classificação deverá ser 62,5% (62,5% = 1,25 × 50%). Se a precisão de chances for de 30%, a precisão de classificação deverá ser 37,5% (37,5% = 1,25 × 30%).

Esse critério fornece apenas uma estimativa grosseira do nível aceitável de precisão preditiva. O critério é fácil de aplicar com grupos de mesmo tamanho. Com grupos de tamanhos desiguais, um limite superior é alcançado quando o modelo de chance máxima é usado para determinar a precisão de chances. No entanto, isso não representa um grande problema, pois sob a maioria das circunstâncias o modelo de chance máxima não seria usado com grupos de tamanhos distintos.

Razões de sucesso geral versus específicas de grupos. Até este ponto, nos concentramos no cálculo da razão de sucesso geral em todos os grupos avaliando a precisão preditiva de uma análise discriminante. O pesquisador também deve estar preocupado com a razão de sucesso (percentual corretamente classificado) para cada grupo separado. Se você se concentrar somente na razão de sucesso geral, é possível que um ou mais grupos, particularmente os menores, possam ter razões de sucesso inaceitáveis enquanto a razão de sucesso geral é aceitável. O pesquisador deve calcular a razão de sucesso de cada grupo e avaliar se a análise discriminante fornece níveis adequados de precisão preditiva tanto no nível geral quanto para cada grupo.

Medidas com base estatística de precisão de classificação relacionada a chances* Um teste estatístico do poder discriminatório da matriz de classificação quando comparada com um modelo de chances é a **estatística *Q* de Press**. Essa medida simples compara o número de classificações corretas com o tamanho da amostra total e o número de grupos. O valor calculado é então comparado com um valor crítico (o valor qui-quadrado para um grau de liberdade no nível de confiança desejado). Se ele excede este valor crítico, então a matriz de classificação pode ser considerada estatisticamente melhor do que as chances. A estatística *Q* é calculada pela seguinte fórmula:

$$Q \text{ de Press} = \frac{[N-(nK)]^2}{N(K-1)}$$

onde

N = tamanho da amostra total
n = número de observações corretamente classificadas
K = número de grupos

Por exemplo, na Tabela 5-4, a estatística *Q* seria baseada em uma amostra total de $N = 50$, $n = 42$ observações corretamente classificadas, e $K = 2$ grupos. A estatística calculada seria:

$$Q \text{ de Press} = \frac{[50-(42\times 2)]^2}{50(2-1)} = 23{,}12$$

O valor crítico em um nível de significância de 0,01 é 6,63. Assim, concluiríamos que, no exemplo, as previsões seriam significantemente melhores do que chances, as quais teriam uma taxa de classificação correta de 50%.

Esse teste simples é sensível ao tamanho da amostra; amostras grandes são mais prováveis de mostrar significância do que amostras pequenas da mesma taxa de classificação.

Por exemplo, se o tamanho da amostra é aumentado para 100 no exemplo e a taxa de classificação permanece em 84%, a estatística *Q* aumenta para 46,24. Se o tamanho da amostra sobe para 200, mas mantém a taxa de classificação em 84%, a estatística *Q* novamente aumenta para 92,48%. Mas se a amostra for apenas 20 e a taxa de classificação incorreta** for ainda de 84% (17 previsões corretas), a estatística *Q* seria de somente 9,8. Ou seja, examine a estatística *Q* à luz do tamanho amostral, pois aumentos no tamanho da amostra fazem subir a estatística *Q* ainda que seja para a mesma taxa de classificação geral.

* N. de R. T.: A palavra "chance" também poderia ser traduzida como "acaso".

** N. de R. T.: A frase correta seria "taxa de classificação correta".

Porém, é necessário cuidado nas conclusões baseadas apenas nessa estatística, pois à medida que a amostra fica maior, uma taxa de classificação menor ainda será considerada significante.

Diagnóstico por casos

O meio final de avaliar o ajuste de modelo é examinar os resultados preditivos em uma base de casos. Semelhante à análise de resíduos em regressão múltipla, o objetivo é entender quais observações (1) foram mal classificadas e (2) não são representativas dos demais membros do grupo. Apesar de a matriz de classificação fornecer precisão de classificação geral, ela não detalha os resultados individuais. Além disso, mesmo que possamos denotar quais casos são correta ou incorretamente classificados, ainda precisamos de uma medida da similaridade de uma observação com o restante do grupo.

Má classificação de casos individuais

Quando se analisam resíduos de uma análise de regressão múltipla, uma decisão importante envolve estabelecer o nível de resíduo considerado substancial e merecedor de atenção. Em análise discriminante, essa questão é mais simples, porque uma observação é ou correta, ou incorretamente classificada. Todos os programas de computador fornecem informação que identifica quais casos são mal classificados e para quais grupos eles foram mal classificados. O pesquisador pode identificar não apenas aqueles casos com erros de classificação, mas uma representação direta do tipo de má classificação.

Análise de casos mal classificados

O propósito de identificar e analisar as observações mal classificadas é identificar quaisquer características dessas observações que pudessem ser incorporadas à análise discriminante para melhorar a precisão preditiva. Essa análise pode assumir a forma de se estabelecer o perfil de casos mal classificados tanto nas variáveis independentes quanto em outras variáveis não incluídas no modelo.

O perfil das variáveis independentes. Examinar esses casos nas variáveis independentes pode identificar tendências não-lineares ou outras relações ou atributos que conduziram à má classificação. Várias técnicas são particularmente adequadas em análise discriminante:

- Uma representação gráfica das observações é talvez a abordagem mais simples e efetiva para examinar as características de observações, especialmente as mal classificadas. A abordagem mais comum é fazer o gráfico das observações com base em seus escores *Z* discriminantes e mostrar a sobreposição entre grupos e os casos mal classificados. Se duas ou mais funções são mantidas, os pontos de corte ótimo também podem ser representados graficamente para fornecer aquilo que é conhecido como um **mapa territorial**, que exibe as regiões correspondentes para cada grupo.

- Representar graficamente as observações individuais com os centróides dos grupos, como anteriormente discutido, mostra não apenas as características gerais dos grupos via centróides, mas também a variação nos membros nos grupos. Isso é análogo às áreas definidas no exemplo de três grupos no começo deste capítulo, em que escores de corte em ambas as funções definiam áreas correspondentes às previsões de classificação para cada grupo.
- Uma avaliação empírica direta da similaridade de uma observação com os membros do outro grupo pode ser feita calculando-se a distância D^2 de Mahalanobis da observação ao centróide do grupo. Com base no conjunto de variáveis independentes, observações mais próximas ao centróide têm um D^2 de Mahalanobis menor e são consideradas mais representativas do grupo do que as mais afastadas.
- No entanto, a medida empírica deve ser combinada com uma análise gráfica, pois apesar de um grande D^2 de Mahalanobis indicar observações que são bastante diferentes dos centróides de grupo, isso nem sempre indica má classificação. Por exemplo, em uma situação de dois grupos, um membro do grupo A pode ter uma grande distância D^2 de Mahalanobis, indicando que ele é menos representativo do grupo. Contudo, se essa distância está afastada do centróide do grupo B, então realmente aumentam as chances de classificação correta, mesmo que ele seja menos representativo do grupo. Uma menor distância que coloca uma observação entre os dois centróides provavelmente teria uma menor probabilidade de classificação correta, mesmo que ela esteja mais próxima ao centróide de seu grupo do que na situação anterior.

Apesar de não existir qualquer análise pré-especificada, como na regressão múltipla, o pesquisador é encorajado a avaliar esses casos mal classificados de diversos pontos de vista, na tentativa de descobrir as características únicas que eles têm em comparação com os outros membros do seu grupo.

Perfil de variáveis não presentes na análise. O exame de outras variáveis quanto às suas diferenças nos casos mal classificados seria o primeiro passo para sua possível inclusão na análise discriminante. Muitas vezes, variáveis que discriminam apenas em um conjunto menor de casos não são identificadas no primeiro conjunto de análises, mas se tornam mais evidentes na análise de casos mal classificados. O pesquisador é encorajado a rever as áreas de suporte conceitual para identificar novas possíveis variáveis que possam se relacionar unicamente com os casos mal classificados e aumentar a precisão preditiva geral.

Resumo

O estágio de estimação e avaliação tem várias semelhanças com as outras técnicas de dependência, permitindo um processo de estimação direta ou *stepwise* e uma análise da precisão preditiva geral e de casos. O pesquisador deve dedicar considerável atenção a essas questões para evitar o uso de um modelo de análise discriminante fundamentalmente errado.

REGRAS PRÁTICAS 5-3

Avaliação do ajuste de modelo e precisão preditiva

- A matriz de classificação e a razão de sucesso substituem R^2 como a medida de ajuste de modelo:
  - Avalie a razão de sucesso geral e por grupo
  - Se as amostras de estimação e análise excederem 100 casos e cada grupo exceder 20 casos, derive padrões separados para cada amostra; caso contrário, derive um único padrão a partir da amostra geral
- Critérios múltiplos são usados para comparação com a razão de sucesso:
  - O critério de chance máxima para avaliação da razão de sucesso é o mais conservador, dando a mais elevada base para exceder
  - Seja cuidadoso no uso do critério de chance máxima em situações com amostras gerais menores que 100 e/ou grupos com menos de 20
  - O critério de chance proporcional considera todos os grupos no estabelecimento do padrão de comparação e é o mais popular
  - A verdadeira precisão preditiva (razão de sucesso) deve exceder qualquer valor de critério em pelo menos 25%
- Analise as observações mal classificadas gráfica (mapa territorial) e empiricamente (D^2 de Mahalanobis)

ESTÁGIO 5: INTERPRETAÇÃO DOS RESULTADOS

Se a função discriminante é estatisticamente significante e a precisão de classificação é aceitável, o pesquisador deve se concentrar em fazer interpretações substanciais das descobertas. Esse processo envolve o exame das funções discriminantes para determinar a importância relativa de cada variável independente na discriminação entre os grupos. Três métodos para determinar a importância relativa foram propostos:

1. Pesos discriminantes padronizados
2. Cargas discriminantes (correlações de estrutura)
3. Valores *F* parciais

Pesos discriminantes

A abordagem tradicional para interpretar funções discriminantes examina o sinal e a magnitude do **peso discriminante** padronizado (às vezes chamado de **coeficiente discriminante**) designado para cada variável ao se computarem as funções discriminantes. Quando o sinal é ignorado, cada peso representa a contribuição relativa de sua variável associada àquela função. As variáveis independentes com pesos relativamente maiores contribuem mais para o poder discriminatório da função do que as variáveis com pesos menores. O sinal indica

apenas que a variável tem uma contribuição positiva ou negativa [4].

A interpretação de pesos discriminantes é análoga à interpretação de pesos beta em análise de regressão e está, portanto, sujeita às mesmas críticas. Por exemplo, um peso pequeno pode indicar que sua variável correspondente é irrelevante na determinação de uma relação, ou que ela tenha sido deixada de lado na relação por causa de um elevado grau de multicolinearidade. Um outro problema do uso de pesos discriminantes é que eles estão sujeitos a considerável instabilidade. Esses problemas sugerem cuidado ao se usarem pesos para interpretar os resultados da análise discriminante.

Cargas discriminantes

As **cargas discriminantes**, às vezes chamadas de **correlações de estrutura**, são cada vez mais usadas como uma base para interpretação, por conta das deficiências na utilização de pesos. Medindo a correlação linear simples entre cada variável independente e a função discriminante, as cargas discriminantes refletem a variância que as variáveis independentes compartilham com a função discriminante. Em relação a isso, elas podem ser interpretadas como cargas fatoriais na avaliação da contribuição relativa de cada variável independente para a função discriminante. (O Capítulo 3 discute melhor a interpretação de cargas fatoriais.)

Uma característica ímpar de cargas é que elas podem ser calculadas para todas as variáveis, sejam elas usadas na estimação da função discriminante ou não. Este aspecto é particularmente útil quando um processo de estimação *stepwise* é empregado e algumas variáveis não são incluídas na função discriminante. Em vez de não se ter forma alguma de compreender seu impacto relativo, as cargas fornecem um efeito relativo de cada variável em uma medida comum.

Com as cargas, a questão principal é: Quais valores as cargas devem assumir para serem consideradas substantivas discriminadoras dignas de nota? Tanto em análise discriminante simultânea quanto *stepwise*, variáveis que exibem uma carga de ± 0,40 ou mais são consideradas substantivas. Com procedimentos *stepwise*, tal determinação é suplementada, pois a técnica evita que variáveis não-significantes entrem na função. Porém, multicolinearidade e outros fatores podem evitar uma variável na equação, o que não significa necessariamente que ela não tenha um efeito substancial.

As cargas discriminantes (assim como os pesos) podem estar sujeitas à instabilidade. As cargas são consideradas relativamente mais válidas do que os pesos como um meio de interpretação do poder discriminatório de variáveis independentes por causa de sua natureza correlacional. O pesquisador ainda deve ser cuidadoso ao usar cargas para interpretar funções discriminantes.

Valores *F* parciais

Como anteriormente discutido, duas abordagens computacionais – simultânea e *stepwise* – podem ser utilizadas para determinar funções discriminantes. Quando o método *stepwise* é selecionado, um meio adicional de interpretar o poder discriminatório relativo das variáveis independentes está disponível pelo uso de valores *F* parciais. Isso é obtido examinando-se os tamanhos absolutos dos valores *F* significantes e ordenando-os. Valores *F* grandes indicam maior poder discriminatório. Na prática, as ordenações que usam a abordagem dos valores *F* são iguais à ordenação determinada a partir do uso de pesos discriminantes, mas os valores *F* indicam o nível associado de significância para cada variável.

Interpretação de duas ou mais funções

Quando há duas ou mais funções discriminantes significantes, temos problemas adicionais de interpretação. Primeiro, podemos simplificar os pesos ou cargas discriminantes para facilitar a determinação do perfil de cada função? Segundo, como representamos o impacto de cada variável nas funções? Esses problemas ocorrem tanto na medida dos efeitos discriminantes totais das funções quanto na avaliação do papel de cada variável no perfil de cada função separadamente. Tratamos dessas duas questões introduzindo os conceitos de rotação das funções, o índice de potência, e representações de vetores expandidos.

Rotação das funções discriminantes

Depois que as funções discriminantes foram desenvolvidas, elas podem ser rotacionadas para redistribuir a variância. (O conceito é melhor explicado no Capítulo 3.) Basicamente, a rotação preserva a estrutura original e a confiabilidade da solução discriminante, ao passo que torna as funções muito mais fáceis de interpretar. Na maioria dos casos, a rotação VARIMAX é empregada como a base para a rotação.

Índice de potência

Anteriormente discutimos o uso de pesos padronizados ou cargas discriminantes como medidas da contribuição de uma variável a uma função discriminante. Quando duas ou mais funções são determinadas, contudo, uma medida resumo ou composta é útil para descrever as contribuições de uma variável em *todas* as funções significantes. O **índice de potência** é uma medida relativa entre todas as variáveis que é indicativa do poder discriminante de cada variável [18]. Ele inclui a contribuição de uma variável a uma função discriminante (sua carga discriminante) e a contribuição relativa da função para a solução geral (uma medida relativa entre as funções com base nos autovalores). A composição é simplesmente a soma dos índices de potência individuais em todas as funções discriminantes significan-

tes. A interpretação da medida composta é limitada, contudo, pelo fato de que é útil apenas na representação da posição relativa (como o oposto de uma ordenação) de cada variável, e o valor absoluto não tem qualquer significado real. O índice de potência é calculado por um processo de dois passos:

Passo 1: *Calcular um valor de potência para cada função significante.* No primeiro passo, o poder discriminatório de uma variável, representado pelo quadrado da carga discriminante não-rotacionada, é "ponderado" pela contribuição relativa da função discriminante para a solução geral. Primeiro, a medida do autovalor relativo para cada função discriminante significante é calculada simplesmente como:

$$\text{Autovalor relativo da função discriminante } j = \frac{\text{Autovalor da função discriminante } j}{\text{Soma de autovalores em todas as funções significantes}}$$

O valor potência de cada variável em uma função discriminante é então:

$$\text{Valor potência da variável } i \text{ na função } j = (\text{Carga discriminante}_{ij})^2 \times \text{Autovalor relativo da função } j$$

Passo 2: *Calcular um índice de potência composto em todas as funções significantes.* Uma vez que um valor potência tenha sido calculado para cada função, o índice de potência composto para cada variável é calculado como:

$$\text{Potência composta da variável } i = \text{Soma dos valores de potência da variável } i \text{ em todas as funções discriminantes significantes}$$

O índice de potência agora representa o efeito discriminante total da variável em todas as funções discriminantes significantes. É apenas uma medida relativa, contudo, e seu valor absoluto não tem qualquer significado importante. Uma ilustração de cálculo de índice de potência é fornecida no exemplo para análise discriminante de três grupos.

Disposição gráfica de escores e cargas discriminantes

Para representar diferenças nos grupos nas variáveis preditoras, o pesquisador pode usar dois diferentes tratamentos para representação gráfica. O mapa territorial representa graficamente os casos individuais de funções discriminantes significantes para permitir ao pesquisador uma avaliação da posição relativa de cada observação com base nos escores da função discriminante. A segunda abordagem é representar graficamente as cargas discriminantes para entender o agrupamento relativo e a magnitude de cada carga sobre cada função. Cada abordagem será discutida detalhadamente na próxima seção.

Mapa territorial. O método gráfico mais comum é o mapa territorial, no qual cada observação é impressa em um gráfico com base nos escores Z da função discriminante das observações. Por exemplo, considere que uma análise discriminante de três grupos tem duas funções discriminantes significantes. Um mapa territorial é criado fazendo-se o gráfico dos escores Z discriminantes de cada observação para a primeira função discriminante sobre o eixo X e os escores para a segunda função discriminante sobre o eixo Y. Desse modo, isso fornece diversas perspectivas de análise:

- O gráfico dos membros de cada grupo com diferentes símbolos permite um retrato fácil das diferenças de cada grupo, bem como suas sobreposições um com o outro.
- O gráfico dos centróides de cada grupo fornece uma maneira de avaliar cada membro de grupo relativamente ao seu centróide. Este procedimento é particularmente útil na avaliação da possibilidade de grandes medidas de Mahalanobis D^2 conduzirem a classificações ruins.
- Retas representando os escores de corte também podem ser graficamente representadas, denotando fronteiras que representam os intervalos de escores discriminantes previstos em cada grupo. Quaisquer membros de grupos que estejam fora dessas fronteiras são mal classificados. Denotar os casos mal classificados permite uma avaliação sobre qual função discriminante foi mais responsável pela má classificação, e sobre o grau em que um caso é mal classificado.

Gráfico vetorial de cargas discriminantes. A abordagem gráfica mais simples é representar cargas reais rotacionadas ou não-rotacionadas. A abordagem preferencial seria com cargas rotacionadas. Semelhante ao gráfico de cargas fatoriais (ver Capítulo 3), este método representa o grau em que cada variável é associada com cada função discriminante.

Uma técnica ainda mais precisa, porém, envolve o gráfico de cargas bem como vetores para cada carga e centróide de grupo. Um **vetor** é meramente uma reta desenhada a partir da origem (centro) de um gráfico até as coordenadas das cargas de uma variável particular ou um centróide de grupo. Com a representação de um **vetor expandido**, o comprimento de cada vetor se torna indicativo da importância relativa de cada variável na discriminação entre os grupos. O procedimento gráfico segue em três passos:

1. *Seleção de variáveis:* Todas as variáveis, sejam incluídas no modelo ou não, podem ser graficamente representadas como vetores. Desse modo, a importância de variáveis colineares que não estão incluídas, como em *stepwise*, ainda pode ser retratada.
2. *Expansão de vetores:* As cargas discriminantes de cada variável são expandidas multiplicando-se a carga discriminante (preferencialmente após a rotação) por seu respectivo valor F univariado. Notamos que os vetores apontam para os grupos com a maior média sobre o preditor respectivo e na direção oposta dos grupos com os menores escores médios.
3. *Gráfico dos centróides de grupos:* Os centróides de grupo também são expandidos nesse procedimento, sendo multiplicados pelo valor F aproximado associado a cada função

discriminante. Se as cargas são expandidas, os centróides também devem ser expandidos para representá-los com precisão no mesmo gráfico. Os valores *F* aproximados para cada função discriminante são obtidos pela seguinte fórmula:

$$\text{Valor } F_{\text{Função}_i} = \text{Autovalor}_{\text{Função}_i} \left(\frac{N_{\text{Amostra de estimação}} - NG}{NG - 1} \right)$$

onde

$N_{\text{Amostra de estimação}}$ = tamanho da amostra de estimação

Por exemplo, considere que a amostra de 50 observações tenha sido dividida em três grupos. O multiplicador de cada autovalor seria (50 – 3)/(3 – 1) = 23,5.

Quando completado, o pesquisador dispõe de um retrato do agrupamento de variáveis em cada função discriminante, a magnitude da importância de cada variável (representada pelo comprimento de cada vetor) e o perfil de cada centróide de grupo (mostrado pela proximidade de cada vetor). Apesar de este procedimento dever ser feito manualmente na maioria dos casos, ele dá um retrato completo das cargas discriminantes e dos centróides de grupos. Para mais detalhes sobre esse procedimento, ver Dillon e Goldstein [4].

Qual método interpretativo usar?

Diversos métodos para interpretar a natureza das funções discriminantes foram discutidos, tanto para soluções de uma função quanto de múltiplas. Quais métodos devem ser usados? A abordagem das cargas é mais válida do que o emprego de pesos e deve ser utilizada sempre que possível. O uso de valores *F* parciais e univariados permite ao pesquisador empregar diversas medidas e procurar alguma consistência nas avaliações das variáveis. Se duas ou mais funções são estimadas, então o pesquisador pode utilizar diversas técnicas gráficas e o índice de potência, que ajuda na interpretação da solução multidimensional. O ponto mais básico é que o pesquisador deve usar todos os métodos disponíveis para chegar à interpretação mais precisa.

ESTÁGIO 6: VALIDAÇÃO DOS RESULTADOS

O estágio final de uma análise discriminante envolve a validação dos resultados discriminantes para garantir que os resultados têm validade externa e interna. *Com a propensão da análise discriminante para aumentar a razão de sucesso se avaliada apenas sobre a amostra de análise, a validação é um passo essencial.* Além de validar as razões de sucesso, o pesquisador deve usar o perfil de grupos para garantir que as médias de grupos sejam indicadores válidos do modelo conceitual usado na seleção de variáveis independentes.

Procedimentos de validação

Validação é um passo crítico em qualquer análise discriminante, pois muitas vezes, especialmente com amostras menores, os resultados podem carecer de generalidade (validade externa). A técnica mais comum para estabelecer validade externa é a avaliação de razões de sucesso. Validação pode ocorrer com uma amostra separada (amostra de teste) ou utilizando-se um procedimento que repetidamente processa a amostra de estimação. Validade externa é admitida quando a razão de sucesso da abordagem selecionada excede os padrões de comparação que representam a precisão preditiva esperada pelo acaso (ver discussão anterior).

Utilização de uma amostra de teste

Geralmente, a validação das razões de sucesso é executada criando-se uma amostra de teste, também chamada de **amostra de validação.** O propósito de se utilizar uma amostra de teste para fins de validação é ver o quão bem a função discriminante funciona em uma amostra de observações não usadas para obter a mesma. Este processo envolve o desenvolvimento de uma função discriminante com a amostra de análise e então a sua aplicação na amostra de teste. A justificativa para dividir a amostra total em dois grupos é que um viés ascendente ocorrerá na precisão preditiva da função discriminante se os indivíduos usados no desenvolvimento da matriz de classificação forem os mesmos utilizados para computar a função; ou seja, a precisão de classificação será mais alta do que é válido se ela for aplicada na amostra de estimação.

Outros pesquisadores têm sugerido que uma confiança maior ainda poderia ser depositada na validade da função discriminante seguindo-se esse procedimento diversas vezes [18]. Ao invés de dividir aleatoriamente a amostra total em grupos de análise e de teste uma vez, o pesquisador dividiria aleatoriamente a amostra total em amostras de análise e de teste várias vezes, sempre testando a validade da função discriminante pelo desenvolvimento de uma matriz de classificação e de uma razão de sucesso. Então as diversas razões de sucesso teriam uma média para se obter uma única medida.

Validação cruzada

A técnica de validação cruzada para avaliar validade externa é feita com múltiplos subconjuntos da amostra total [2,4]. A abordagem mais amplamente usada é o método *jackknife*. Validação cruzada é baseada no princípio do "deixe um de fora". O uso mais comum desse método é estimar $k - 1$ amostras, eliminando-se uma observação por vez a partir de uma amostra de k casos. Uma fun-

ção discriminante é calculada para cada subamostra, e em seguida a pertinência a grupo prevista da observação eliminada é feita com a função discriminante estimada sobre os demais casos. Depois que todas as previsões de pertinência a grupo foram feitas, uma por vez, uma matriz de classificação é construída e a razão de sucesso é calculada.

Validação cruzada é muito sensível a amostras pequenas. Orientações sugerem que ela seja usada somente quando o tamanho do grupo menor é pelo menos três vezes o número de variáveis preditoras, e a maioria dos pesquisadores sugere uma proporção de cinco para um [13]. No entanto, validação cruzada pode representar a única técnica de validação possível em casos em que a amostra original é muito pequena para dividir em amostras de análise e de teste, mas ainda excede as orientações já discutidas. Validação cruzada também está se tornando mais amplamente usada à medida que os principais programas de computador a disponibilizam como opção.

Diferenças de perfis de grupos

Uma outra técnica de validação é estabelecer o perfil dos grupos sobre as variáveis independentes para garantir sua correspondência com as bases conceituais usadas na formulação do modelo original. Depois que o pesquisador identifica as variáveis independentes que oferecem a maior contribuição à discriminação entre os grupos, o próximo passo é traçar o perfil das características dos grupos com base nas médias dos mesmos. Esse perfil permite ao pesquisador compreender o caráter de cada grupo de acordo com as variáveis preditoras.

REGRAS PRÁTICAS 5-4

Interpretação e validação de funções discriminantes

- Cargas discriminantes são o método preferido para avaliar a contribuição de cada variável em uma função discriminante, pois elas são:
  - Uma medida padronizada de importância (variando de 0 a 1)
  - Disponíveis para todas as variáveis independentes, sejam usadas no processo de estimação ou não
  - Não afetadas por multicolinearidade
- Cargas excedendo ± 0,40 são consideradas substantivas para fins de interpretação
- No caso de mais de uma função discriminante, certifique-se de:
  - Usar cargas rotacionadas
  - Avaliar a contribuição de cada variável em todas as funções com o índice de potência
- A função discriminante deve ser validada com a amostra de teste ou um dos procedimentos "deixe um de fora"

Por exemplo, olhando os dados da pesquisa da Kitchen-Aid apresentados na Tabela 5-1, percebemos que a avaliação média de "durabilidade" para o grupo "compraria" é 7,4, enquanto a avaliação média comparável de "durabilidade" para o grupo "não compraria" é de 3,2. Assim, um perfil desses dois grupos mostra que o grupo "compraria" avalia a durabilidade percebida do novo produto bem mais do que o grupo "não compraria".

Outra abordagem é estabelecer o perfil de grupos em um conjunto separado de variáveis que deve espelhar as diferenças observadas de grupos. Esse perfil separado fornece uma avaliação de validade externa, de modo que os grupos variam tanto na(s) variável(eis) independente(s) quanto no conjunto de variáveis associadas. Essa técnica é semelhante, em caráter, à validação de agrupamentos obtidos descrita no Capítulo 8.

UM EXEMPLO ILUSTRATIVO DE DOIS GRUPOS

Para ilustrar a aplicação da análise discriminante de dois grupos, usamos variáveis obtidas da base de dados HBAT introduzida no Capítulo 1. Esse exemplo examina cada um dos seis estágios do processo de construção de modelo para um problema de pesquisa particularmente adequado à análise discriminante múltipla.

Estágio 1: Objetivos da análise discriminante

Você lembra que uma das características de cliente obtida pela HBAT em sua pesquisa foi uma variável categórica (X_4) que indicava a região na qual a empresa estava localizada: EUA/América do Norte ou fora. A equipe administrativa da HBAT está interessada em quaisquer diferenças de percepções entre aqueles clientes localizados e servidos por sua equipe de venda nos EUA versus aqueles fora dos EUA e que são servidos principalmente por distribuidores independentes. A despeito de diferenças encontradas em termos de suporte de vendas devido à natureza da equipe de venda servindo cada área geográfica, a equipe administrativa está interessada em ver se as outras áreas de operação (linha do produto, preço etc.) são vistas de maneira distinta por estes dois conjuntos de clientes. Esta indagação segue a óbvia necessidade por parte da administração de sempre procurar melhor entender seu cliente, neste caso se concentrando em diferenças que podem ocorrer entre áreas geográficas. Se quaisquer percepções de HBAT forem notadas como diferindo significativamente entre firmas nessas duas regiões, a companhia será então capaz de desen-

volver estratégias para remediar quaisquer deficiências percebidas e desenvolver estratégias diferenciadas para acomodar as percepções distintas.

Para tanto, a análise discriminante foi selecionada para identificar aquelas percepções da HBAT que melhor diferenciam as empresas em cada região geográfica.

Estágio 2: Projeto de pesquisa para análise discriminante

O estágio de projeto de pesquisa se concentra em três questões-chave: selecionar variáveis dependente e independentes, avaliar a adequação do tamanho da amostra para a análise planejada, e dividir a amostra para fins de validação.

Seleção de variáveis dependente e independentes

A análise discriminante requer uma única medida dependente não-métrica e uma ou mais medidas independentes métricas que são afetadas para fornecer diferenciação entre os grupos baseados na medida dependente.

Como a variável dependente Região (X_4) é uma variável categórica de dois grupos, a análise discriminante é a técnica apropriada. O levantamento coletou percepções da HBAT que agora podem ser usadas para distinguir entre os dois grupos de firmas. A análise discriminante usa como variáveis independentes as 13 variáveis de percepção a partir do banco de dados (X_6 a X_{18}) para discriminar entre firmas em cada área geográfica.

Tamanho da amostra

Dado o tamanho relativamente pequeno da amostra HBAT (100 observações), questões como tamanho amostral são particularmente importantes, especialmente a divisão da amostra em amostras de teste e de análise (ver discussão na próxima seção).

A amostra de 100 observações, quando particionada em amostras de análise e de teste de 60 e 40 respectivamente, mal atende à proporção mínima de 5 para 1 de observações para variáveis independentes (60 observações para 13 variáveis independentes em potencial) sugerida para a amostra de análise. Apesar de essa proporção crescer para quase 8 para 1 se a amostra não for dividida, considera-se mais importante validar os resultados do que aumentar o número de observações na amostra de análise.

Os dois grupos de 26 e 34 na amostra de estimação também excedem o tamanho mínimo de 20 observações por grupo. Finalmente, os dois grupos são suficientemente comparáveis em tamanho para não impactar adversamente os processos de estimação ou de classificação.

Divisão da amostra

A discussão anterior enfatizou a necessidade de validar a função discriminante dividindo a amostra em duas partes, uma usada para estimação e a outra para validação. Em qualquer momento em que uma amostra de teste é empregada, o pesquisador deve garantir que os tamanhos de amostra resultantes sejam suficientes para embasar o número de preditores incluídos na análise.

A base de dados HBAT tem 100 observações; foi decidido que uma amostra de teste de 40 observações seria suficiente para fins de validação. Essa partição deixaria ainda 60 observações para a estimação da função discriminante. Além disso, os tamanhos relativos de grupos na amostra de estimação (26 e 34 nos dois grupos) permitiriam a estimação sem complicações devidas a diferenças consideráveis de tamanhos de grupos.

É importante garantir aleatoriedade na seleção da amostra de validação, de modo que qualquer ordenação das observações não afete os processos de estimação e de validação.

Estágio 3: Suposições da análise discriminante

As principais suposições inerentes à análise discriminante envolvem a formação da variável estatística ou função discriminante (normalidade, linearidade e multicolinearidade) e a estimação da função discriminante (matrizes de variância e covariância iguais). Como examinar as variáveis independentes quanto à normalidade, linearidade e multicolinearidade é explicado no Capítulo 2. Para fins de nossa ilustração da análise discriminante, essas suposições são atendidas em níveis aceitáveis.

A maioria dos programas estatísticos tem um ou mais teste(s) estatístico(s) para a suposição de matrizes de covariância ou dispersão iguais abordada no Capítulo 2. O mais comum é o teste M de Box (para mais detalhes, ver Capítulo 2).

Neste exemplo de dois grupos, a significância de diferenças nas matrizes de covariância entre os dois grupos é de 0,011. Mesmo que a significância seja menor que 0,05 (nesse teste o pesquisador procura por valores acima do nível desejado de significância), a sensibilidade do teste a outros fatores que não sejam apenas diferenças de covariância (p.ex., normalidade das variáveis e tamanho crescente da amostra) faz desse um nível aceitável.

Nenhuma ação corretiva adicional faz-se necessária antes que a estimação da função discriminante possa ser realizada.

Estágio 4: Estimação do modelo discriminante e avaliação do ajuste geral

O pesquisador tem a escolha de duas técnicas de estimação (simultânea versus *stepwise*) para determinar as variáveis independentes incluídas na função discriminante. Uma vez que a técnica de estimação é escolhida, o processo determina a composição da função discriminante sujeita à exigência de significância estatística especificada pelo pesquisador.

O principal objetivo dessa análise é identificar o conjunto de variáveis independentes (percepções HBAT) que diferencia ao máximo entre os dois grupos de clientes. Se o conjunto de variáveis de percepções fosse menor ou a meta fosse simplesmente determinar as capacidades discriminantes do conjunto inteiro de variáveis de percepção, sem se preocupar com o impacto de qualquer percepção individual, então a abordagem simultânea de inclusão de todas as variáveis diretamente na função discriminante seria empregada. Mas neste caso, mesmo com o conhecimento de multicolinearidade entre as variáveis de percepção vista no desempenho da análise fatorial (ver Capítulo 3), a abordagem *stepwise* é considerada mais adequada. Devemos observar, porém, que multicolinearidade pode impactar sobre quais variáveis entram na função discriminante e assim exigir particular atenção no processo de interpretação.

Avaliação de diferenças de grupos

Iniciemos nossa avaliação da análise discriminante de dois grupos examinando a Tabela 5-5, que mostra as médias de grupos para cada uma das variáveis independentes, com base nas 60 observações que constituem a amostra de análise.

Para identificar quais das cinco variáveis, mais alguma das demais, melhor discrimina entre os grupos, devemos estimar a função discriminante.

Ao estabelecer o perfil dos dois grupos, podemos primeiramente identificar cinco variáveis com as maiores diferenças nas médias de grupo (X_6, X_{11}, X_{12}, X_{13}, e X_{17}). A Tabela 5-5 também exibe o lambda de Wilks e a ANOVA univariada utilizada para avaliar a significância entre médias das variáveis independentes para os dois grupos. Esses testes indicam que as cinco variáveis de percepção são também as únicas com diferenças univariadas significantes entre os dois grupos. Finalmente, os valores D^2 de Mahalanobis mínimos são também dados. Este valor é importante porque ele é a medida usada para selecionar variáveis para entrada no processo de estimação *stepwise*. Como apenas dois grupos estão envolvidos, o maior valor D^2 tem também a diferença entre grupos mais significante (note que o mesmo fato não ocorre necessariamente com três ou mais grupos, nos quais grandes diferenças entre dois grupos quaisquer podem não resultar nas maiores diferenças gerais em todos os grupos, como será mostrado no exemplo de três grupos).

O exame das diferenças de grupos leva à identificação de cinco variáveis de percepção (X_6, X_{11}, X_{12}, X_{13} e X_{17}) como o conjunto mais lógico de candidatos a entrarem na análise discriminante. Essa considerável redução a partir do conjunto maior de 13 variáveis de percepção reforça a decisão de se usar um processo de estimação *stepwise*.

Estimação da função discriminante

O procedimento *stepwise* começa com todas as variáveis excluídas do modelo e então seleciona a variável que:

1. Mostra diferenças estatisticamente significantes nos grupos (0,05 ou menos exigido para entrada)
2. Dá a maior distância de Mahalanobis (D^2) entre os grupos

Este processo continua a incluir variáveis na função discriminante desde que elas forneçam discriminação adicional estatisticamente significante entre os grupos além daquelas diferenças já explicadas pelas variáveis na função discriminante.Esta técnica é semelhante ao processo *stepwise* em regressão múltipla (ver Capítulo 4), que adiciona variáveis com aumentos significantes na variância explicada da variável dependente. Além disso, em casos nos quais duas ou mais variáveis entram no modelo, as variáveis já presentes são avaliadas para possível remoção. Uma variável pode ser removida se existir elevada multicolinearidade entre ela e as demais variáveis independentes incluídas, de modo que sua significância fica abaixo do nível para remoção (0,10).

Estimação *stepwise*: adição da primeira variável X_{13}. A partir de nossa revisão de diferenças de grupos, percebemos que X_{13} tinha a maior diferença significante entre grupos e o maior D^2 de Mahalanobis (ver Tabela 5-5). Logo, X_{13} entra como a primeira variável no procedimento *stepwise* (ver Tabela 5-6). Como apenas uma variável entra no modelo discriminante neste momento, os níveis de significância e as medidas de diferenças de grupos coincidem com aqueles dos testes univariados.

Depois que X_{13} entra no modelo, as demais variáveis são avaliadas com base em suas habilidades discriminantes incrementais (diferenças de médias de grupos depois

(*Continua*)

TABELA 5-5 Estatísticas descritivas de grupo e testes de igualdade para a amostra de estimação na análise discriminante de dois grupos

		Médias de grupos da variável dependente: X_4 Região		*Teste de igualdade de médias de grupos**			*D^2 de Mahalanobis mínimo*	
	Variáveis independentes	**Grupo 0: EUA/América do Norte (*n* = 26)**	**Grupo 1: Fora da América do Norte (*n* = 34)**	**Lambda de Wilks**	**Valor *F***	**Significância**	**D^2 mínimo**	**Entre grupos**
X_6	Qualidade do produto	8,527	7,297	0,801	14,387	0,000	0,976	0 e 1
X_7	Atividades de Comércio eletrônico	3,388	3,626	0,966	2,054	0,157	0,139	0 e 1
X_8	Suporte técnico	5,569	5,050	0,973	1,598	0,211	0,108	0 e 1
X_9	Solução de reclamação	5,577	5,253	0,986	0,849	0,361	0,058	0 e 1
X_{10}	Anúncio	3,727	3,979	0,987	0,775	0,382	0,053	0 e 1
X_{11}	Linha do produto	6,785	5,274	0,695	25,500	0,000	1,731	0 e 1
X_{12}	Imagem da equipe de venda	4,427	5,238	0,856	9,733	0,003	0,661	0 e 1
X_{13}	Preço competitivo	5,600	7,418	0,645	31,992	0,000	2,171	0 e 1
X_{14}	Garantia e reclamações	6,050	5,918	0,992	0,453	0,503	0,031	0 e 1
X_{15}	Novos produtos	4,954	5,276	0,990	0,600	0,442	0,041	0 e 1
X_{16}	Encomenda e cobrança	4,231	4,153	0,999	0,087	0,769	0,006	0 e 1
X_{17}	Flexibilidade de preço	3,631	4,932	0,647	31,699	0,000	2,152	0 e 1
X_{18}	Velocidade de entrega	3,873	3,794	0,997	0,152	0,698	0,010	0 e 1

* Lambda de Wilks (estatística *U*) e razão *F* univariada com 1 e 58 graus de liberdade.

(*Continuação*)

que a variância associada com X_{13} é removida). Novamente, variáveis com níveis de significância maiores que 0,05 são eliminadas de consideração para entrada no próximo passo.

O exame das diferenças univariadas mostradas na Tabela 5-5 identifica X_{17} (Flexibilidade de preço) como a variável com a segunda maior diferença. No entanto, o processo *stepwise* não utiliza esses resultados univariados quando a função discriminante tem uma ou mais variáveis. Ele calcula os valores D^2 e os testes de significância estatística de diferenças de grupos depois que o efeito das variáveis nos modelos é removido (neste caso apenas X_{13} está no modelo).

Como mostrado na última parte da Tabela 5-6, três variáveis (X_6, X_{11} e X_{17}) claramente atendem ao critério de nível de significância de 0,05 para consideração no próximo estágio. X_{17} permanece como o próximo melhor candidato a entrar no modelo porque ela tem o maior D^2 de Mahalanobis (4,300) e o maior valor F a entrar. Não obstante, outras variáveis (p.ex., X_{11}) têm substanciais reduções em seu nível de significância e no D^2 de Mahalanobis em relação ao que se mostra na Tabela 5-5 devido à variável única no modelo (X_{13}).

Estimação *stepwise*: adição da segunda variável X_{17}. No passo 2 (ver Tabela 5-7), X_{17} entra no modelo, conforme esperado. O modelo geral é significante (F = 31,129) e melhora a discriminação entre grupos, como evidenciado pela diminuição no lambda de Wilks de 0,645 para 0,478. Além disso, o poder discriminante de ambas as variáveis incluídas nesse ponto é também estatisticamente significante (valores F de 20,113 para X_{13} e 19,863 para X_{17}). Com ambas as variáveis estatisticamente significantes, o procedimento se dirige para o exame das variáveis fora da equação na busca de potenciais candidatos para inclusão na função discriminante com base em sua discriminação incremental entre os grupos.

X_{11} é a próxima variável a atender às exigências para inclusão, mas seu nível de significância e sua habilidade discriminante foram reduzidos substancialmente por conta da multicolinearidade com X_{13} e X_{17} já na função discriminante. Mais notável ainda é o considerável aumento no D^2 de Mahalanobis em relação aos resultados univariados nos quais cada variável é considerada separadamente. No caso de $X_{11,}$ o valor D^2 mínimo aumenta de 1,731 (ver Tabela 5-5) para 5,045 (Tabela 5-7), o que indica um espalhamento e uma separação dos grupos por conta de X_{13} e X_{17} já na função discriminante. Note que X_{18} é quase idêntica em poder discriminante remanescente, mas X_{11} entrará no terceiro passo devido à sua pequena vantagem.

Estimação *stepwise*: adição de uma terceira variável X_{11}. A Tabela 5-8 revê os resultados do terceiro passo do processo *stepwise*, onde X_{11} entra na função discriminante. Os resultados gerais ainda são estatisticamente significantes e continuam a melhorar na discriminação, como evidenciado pela diminuição no valor lambda de

(*Continua*)

TABELA 5-6 Resultados do passo 1 da análise discriminante *stepwise* de dois grupos

Ajuste geral do modelo

	Valor	Valor *F*	Graus de liberdade	Significância
Lambda de Wilks	0,645	31,992	1,58	0,000

Variáveis adicionadas/removidas no passo 1

		F		
Variável adicionada	D^2 mínimo	Valor	Significância	Entre grupos
X_{13} Preços competitivos	2,171	31,992	0,000	0 e 1

Nota: Em cada passo, a variável que maximiza a distância de Mahalanobis entre os dois grupos mais próximos é adicionada.

Variáveis na análise após o passo 1

Variável	Tolerância	*F* para remover	D^2	Entre grupos
X_{13} Preços competitivos	1,000	31,992		

Variáveis fora da análise após o passo 1

Variável	Tolerância	Tolerância mínima	*F* para entrar	D^2 mínimo	Entre grupos
X_6 Qualidade de produto	0,965	0,965	4,926	2,699	0 e 1
X_7 Atividades de comércio eletrônico	0,917	0,917	0,026	2,174	0 e 1
X_8 Suporte técnico	0,966	0,966	0,033	2,175	0 e 1
X_9 Solução de reclamação	0,844	0,844	1,292	2,310	0 e 1
X_{10} Anúncio	0,992	0,992	0,088	2,181	0 e 1
X_{11} Linha de produto	0,849	0,849	6,076	2,822	0 e 1
X_{12} Imagem da equipe de venda	0,987	0,987	3,949	2,595	0 e 1
X_{14} Garantia e reclamações	0,918	0,918	0,617	2,237	0 e 1
X_{15} Novos produtos	1,000	1,000	0,455	2,220	0 e 1
X_{16} Encomenda e cobrança	0,836	0,836	3,022	2,495	0 e 1
X_{17} Flexibilidade de preço	1,000	1,000	19,863	4,300	0 e 1
X_{18} Velocidade de entrega	0,910	0,910	1,196	2,300	0 e 1

Teste de significância de diferenças de grupos após o passo 1[a]

		EUA/América do Norte
Fora da América do Norte	*F*	31,992
	Sig.	0,000

[a]1,58 graus de liberdade

(*Continuação*)

Wilks (de 0,478 para 0,438). Note, porém, que a queda foi muito menor do que aquela encontrada quando a segunda variável (X_{17}) foi adicionada à função discriminante. Com X_{13}, X_{17} e X_{11} estatisticamente significantes, o procedimento se dirige para a identificação de candidatos remanescentes para inclusão.

Como visto na última parte da Tabela 5-8, nenhuma das 10 variáveis independentes que sobraram passam pelo critério de entrada de significância estatística de 0,05. Depois que X_{11} entrou na equação, as duas variáveis remanescentes que tinham diferenças univariadas significantes nos grupos (X_6 e X_{12}) apresentam um poder discriminatório adicional relativamente pequeno e não atendem ao critério de entrada. Assim, o processo de estimação pára com as três variáveis (X_{13}, X_{17} e X_{11}) constituindo a função discriminante.

Resumo do processo de estimação *stepwise*. A Tabela 5-9 fornece os resultados gerais da análise discriminante *stepwise* depois que todas as variáveis significantes foram incluídas na estimação da função discriminante. Essa tabela resumo descreve as três variáveis (X_{11}, X_{13} e X_{17}) que são discriminadores significantes com base em seus lambda de Wilks e nos valores mínimos de D^2 de Mahalanobis.

Diversos resultados distintos são dados abordando tanto o ajuste geral do modelo quanto o impacto de variáveis específicas.

(*Continua*)

TABELA 5-7 Resultados do passo 2 da análise discriminante *stepwise* de dois grupos

Ajuste geral do modelo

	Valor	Valor *F*	Graus de liberdade	Significância
Lambda de Wilks	0,478	31,129	2,57	0,000

Variáveis adicionadas/removidas no passo 2

		F		
Variável adicionada	D^2 mínimo	Valor	Significância	Entre grupos
X_{13} Flexibilidade de preço	4,300	31,129	0,000	0 e 1

Nota: Em cada passo, a variável que maximiza a distância de Mahalanobis entre os dois grupos mais próximos é adicionada.

Variáveis na análise após o passo 2

Variável	Tolerância	F para remover	D^2	Entre grupos
X_{13} Preços competitivos	1,000	20,113	2,152	0 e 1
X_{17} Flexibilidade de preço	1,000	19,863	2,171	0 e 1

Variáveis fora da análise após o passo 2

Variável	Tolerância	Tolerância mínima	*F* para entrar	D^2 mínimo	Entre grupos
X_6 Qualidade de produto	0,884	0,884	0,681	4,400	0 e 1
X_7 Atividades de comércio eletrônico	0,804	0,804	2,486	4,665	0 e 1
X_8 Suporte técnico	0,966	0,966	0,052	4,308	0 e 1
X_9 Solução de reclamação	0,610	0,610	1,479	4,517	0 e 1
X_{10} Anúncio	0,901	0,901	0,881	4,429	0 e 1
X_{11} Linha de produto	0,848	0,848	5,068	5,045	0 e 1
X_{12} Imagem da equipe de venda	0,944	0,944	0,849	4,425	0 e 1
X_{14} Garantia e reclamações	0,916	0,916	0,759	4,411	0 e 1
X_{15} Novos produtos	0,986	0,986	0,017	4,302	0 e 1
X_{16} Encomenda e cobrança	0,625	0,625	0,245	4,336	0 e 1
X_{18} Velocidade de entrega	0,519	0,519	4,261	4,927	0 e 1

Teste de significância de diferenças de grupos após o passo 2[a]

		EUA/América do Norte
Fora da América do Norte	*F*	32,129
	Sig.	0,000

[a]2,57 graus de liberdade

(*Continuação*)

- As medidas multivariadas de ajuste geral do modelo são relatadas sob a legenda "Funções discriminantes canônicas". Observe que a função discriminante é altamente significante (0,000) e retrata uma correlação canônica de 0,749. Interpretamos essa correlação elevando-a ao quadrado $(0,749)^2 = 0,561$. Logo, 56,1% da variância na variável dependente (X_4) pode ser explicada por este modelo, o qual inclui apenas três variáveis independentes.
- Os coeficientes padronizados da função discriminante são fornecidos, mas são menos preferidos para fins de interpretação do que as cargas discriminantes. Os coeficientes discriminantes não-padronizados são usados para calcular os escores *Z* discriminantes que podem ser empregados na classificação.
- As cargas discriminantes são relatadas sob a legenda "Matriz estrutural" e são ordenadas da maior para a menor em termos de tamanho da carga. As cargas são discutidas depois na fase de interpretação (estágio 5).
- Os coeficientes da função de classificação, também conhecidos como funções discriminantes lineares de Fisher, são utilizados na classificação e discutidos posteriormente.
- Centróides de grupo são também relatados, e eles representam a média dos escores individuais da função discriminante para cada grupo. Centróides fornecem uma medida resumo da posição relativa de cada grupo nas funções discriminantes. Neste caso, a Tabela 5-9 revela que o centróide de grupo para as firmas nos EUA/América do Norte (grupo 0) é –1,273, enquanto

(*Continua*)

TABELA 5-8 Resultados do passo 3 da análise discriminante *stepwise* de dois grupos

Ajuste geral do modelo

	Valor	Valor *F*	Graus de liberdade	Significância
Lambda de Wilks	0,438	23,923	3, 56	0,000

Variáveis adicionadas/removidas no passo 3

		F		
	D^2 mínimo	Valor	Significância	Entre grupos
X_{11} Linha de produto	5,045	23,923	0,000	0 e 1

Nota: Em cada passo, a variável que maximiza a distância de Mahalanobis entre os dois grupos mais próximos é adicionada.

Variáveis na análise após o passo 3

Variável	Tolerância	*F* para remover	D^2	Entre grupos
X_{13} Preços competitivos	0,849	7,258	4,015	0 e 1
X_{17} Flexibilidade de preço	0,999	18,416	2,822	0 e 1
X_{11} Linha de produto	0,848	5,068	4,300	0 e 1

Variáveis fora da análise após o passo 3

Variável	Tolerância	Tolerância mínima	*F* para entrar	D^2 mínimo	Entre grupos
X_6 Qualidade de produto	0,802	0,769	0,019	5,048	0 e 1
X_7 Atividades de comércio eletrônico	0,801	0,791	2,672	5,482	0 e 1
X_8 Suporte técnico	0,961	0,832	0,004	5,046	0 e 1
X_9 Solução de reclamação	0,233	0,233	0,719	5,163	0 e 1
X_{10} Anúncio	0,900	0,840	0,636	5,149	0 e 1
X_{12} Imagem da equipe de venda	0,931	0,829	1,294	5,257	0 e 1
X_{14} Garantia e reclamações	0,836	0,775	2,318	5,424	0 e 1
X_{15} Novos produtos	0,981	0,844	0,076	5,058	0 e 1
X_{16} Encomenda e cobrança	0,400	0,400	1,025	5,213	0 e 1
X_{18} Velocidade de entrega	0,031	0,031	0,208	5,079	0 e 1

Teste de significância de diferenças de grupos após o passo 3[a]

		EUA/América do Norte
Fora da América do Norte	*F*	23,923
	Sig.	0,000

[a]3,56 graus de liberdade

(*Continuação*)

o centróide para as firmas fora da América do Norte (grupo 1) é 0,973. Para mostrar que a média geral é 0, multiplique o número em cada grupo por seu centróide e some ao resultado (p.ex., $26 \times -1{,}273 + 34 \times 0{,}973 = 0{,}0$).

Os resultados do modelo geral são aceitáveis com base em significância estatística e prática. No entanto, antes de proceder com uma interpretação dos resultados, o pesquisador precisa avaliar a precisão de classificação e examinar os resultados caso a caso.

Avaliação da precisão de classificação

Com o modelo geral estatisticamente significante e explicando 56% da variação entre os grupos (ver a discussão anterior e a Tabela 5-9), passamos para a avaliação de precisão preditiva da função discriminante. Em tal processo devemos completar três tarefas:

1. Calcular o escore de corte, o critério no qual o escore *Z* discriminante de cada observação é julgado para determinar em qual grupo ela deve ser classificada.
2. Classificar cada observação e desenvolver as matrizes de classificação para as amostras de análise e de teste.
3. Avaliar os níveis de precisão preditiva a partir das matrizes de classificação quanto a significância estatística e prática.

Apesar de o exame da amostra de teste e de sua precisão preditiva ser realmente feito no estágio de validação, os resultados são discutidos agora para facilitar a comparação entre as amostras de estimação e de teste.

TABELA 5-9 Estatísticas resumo para análise discriminante de dois grupos

Ajuste geral do modelo: funções discriminantes canônicas

		Percentual de variância						
Função	**Autovalor**	**Função %**	**Cumulativo %**	**Correlação canônica**	**Lambda de Wilks**	**Qui-quadrado**	***df***	**Significância**
1	1,282	100	100	0,749	0,438	46,606	3	0,000

Função discriminante e coeficientes da função de classificação

	Funções discriminantes		*Funções de classificação*	
Variáveis independentes	**Não-padronizado**	**Padronizado**	**Grupo 0: EUA/América do Norte**	**Grupo 1: Fora da América do Norte**
X_{11} Linha de produto	–0,363	–0,417	7,725	6,909
X_{13} Preços competitivos	0,398	0,490	6,456	7,349
X_{17} Flexibilidade de preço	0,749	0,664	4,231	5,912
Constante	–3,752		–52,800	–60,623

Matriz estrutural[a]

Variáveis independentes	**Função 1**
X_{13} Preços competitivos	0,656
X_{17} Flexibilidade de preço	0,653
X_{11} Linha de produto	–0,586
X_7 Atividades de comércio eletrônico*	0,429
X_6 Qualidade de produto*	–0,418
X_{14} Garantia e reclamações*	–0,329
X_{10} Anúncio*	0,238
X_9 Solução de reclamações*	–0,181
X_{12} Imagem da equipe de venda*	0,164
X_{16} Encomenda e cobrança*	–0,149
X_8 Suporte técnico*	–0,136
X_{18} Velocidade de entrega*	–0,060
X_{15} Novos produtos*	0,041

*Variável não usada na análise

Médias de grupos (centróides) de funções discriminantes

X_4 Região	**Função 1**
EUA/América do Norte	–1,273
Fora da América do Norte	0,973

[a]Correlações internas de grupos entre variáveis discriminantes e funções discriminantes canônicas padronizadas ordenadas por tamanho absoluto de correlação na função,

Cálculo do escore de corte. O pesquisador deve primeiramente determinar como as probabilidades *a priori* de classificação são determinadas, ou com base nos tamanhos reais dos grupos (assumindo que eles são representativos da população), ou especificadas pelo pesquisador, sendo que mais freqüentemente são estabelecidas como iguais em uma postura conservadora do processo de classificação.

Nesta amostra de análise de 60 observações, sabemos que a variável dependente consiste em dois grupos, 26 empresas localizadas nos EUA e 34 empresas fora do país. Se não estamos certos de que as proporções da população são representadas pela amostra, então devemos empregar probabilidades iguais. No entanto, como nossa amostra de empresas é aleatoriamente extraída, podemos estar razoavelmente certos de que essa amostra reflete as proporções da população. Logo, essa análise discriminante usa as proporções da amostra para especificar as probabilidades *a priori* para fins de classificação. Tendo especificado as probabilidades *a priori*, o escore de corte ótimo pode ser calculado. Como nesta situação os grupos são considerados representativos, o cálculo se torna uma média ponderada dos dois centróides de grupos:

(*Continua*)

(*Continuação*)

$$Z_{CS} = \frac{N_A Z_B + N_B Z_A}{N_A + N_B} = \frac{(26 \times 0,973) + (34 \times -1,273)}{26 + 34} = -0,2997$$

Substituindo os valores apropriados na fórmula, podemos obter o escore de corte crítico (assumindo custos iguais de má classificação) de $Z_{CS} = -0,2997$.

Classificação de observações e construção de matrizes de classificação. Uma vez que o escore de corte tenha sido calculado, cada observação pode ser classificada comparando seu escore discriminante com o de corte.

O procedimento para classificar empresas com o escore de corte ótimo é o seguinte:

- Classifique uma empresa como sendo do grupo 0 (Estados Unidos/América do Norte) se seu escore discriminante for menor que –0,2997.
- Classifique uma empresa como sendo do grupo 1 (Fora dos Estados Unidos) se seu escore discriminante for maior que –0,2997.

Matrizes de classificação para as observações nas amostras de análise e de validação foram calculadas, e os resultados são exibidos na Tabela 5-10. A amostra de análise tem 86,7% de precisão de previsão, que é ligeiramente maior que a precisão de 85% da amostra de teste, como já antecipado. Além disso, a amostra que passou por validação cruzada conseguiu uma precisão preditiva de 83,3%.

Avaliação da precisão de classificação atingida. Ainda que todas as medidas de precisão de classificação sejam bastante altas, o processo de avaliação requer uma comparação com a precisão de classificação em uma série de medidas baseadas em chances. Essas medidas refletem a melhora do modelo discriminante quando se compara com a classificação de indivíduos sem o uso da função discriminante. Sabendo-se que a amostra geral é de 100 observações e que os grupos de teste/validação são menores do que 20, usaremos a amostra geral para estabelecer os padrões de comparação.

A primeira medida é o critério de chance proporcional, o qual considera que os custos da má classificação são iguais (ou seja, queremos identificar os membros de cada grupo igualmente bem). O critério de chance proporcional é:

$$C_{PRO} = p^2 + (1 - p)^2$$

onde

C_{PRO} = critério de chance proporcional
p = proporção de empresas no grupo 0
$1 - p$ = proporção de empresas no grupo 1

O grupo de clientes localizados nos Estados Unidos (grupo 0) constitui 39,0% da amostra de análise (39/100), com o segundo grupo representando clientes localizados fora dos Estados Unidos (grupo 1) formando os 61,0% restantes. O valor calculado de chance proporcional é de 0,524 ($0,390^2 + 0,610^2 = 0,524$).

O critério de chance máxima é simplesmente o percentual corretamente classificado se todas as observações fossem colocadas no grupo com a maior probabilidade de ocorrência. Ele reflete nosso padrão mais conserva-

TABELA 5-10 Resultados de classificação para análise discriminante de dois grupos

Resultados de classificação[a, b, c]

		Pertinência prevista em grupo		
Amostra	**Grupo real**	**EUA/América do Norte**	**Fora da América do Norte**	**Total**
Amostra de estimação	EUA/América do Norte	25	1	26
		96,2%	3,8%	
	Fora da América do Norte	7	27	34
		20,6%	79,4%	
Amostra de validação cruzada[d]	EUA/América do Norte	24	2	26
		92,3	7,7	
	Fora da América do Norte	8	26	34
		23,5	76,5	
Amostra de teste	EUA/América do Norte	9	4	13
		69,2	30,8	
	Fora da América do Norte	2	25	27
		7,4	92,6	

[a]86,7% dos casos originais selecionados e agrupados (amostra de estimação) corretamente classificados.
[b]85,0% dos casos originais não-selecionados e agrupados (amostra de validação) corretamente classificados.
[c]83,3% dos casos selecionados validados por cruzamento corretamente classificados.
[d]Validação cruzada é feita somente para aqueles casos da análise (amostra de estimação). Em validação cruzada, cada caso é classificado pelas funções derivadas de todos os casos distintos daquele.

dor e assume nenhuma diferença no custo de uma má classificação.

Como o grupo 1 (clientes fora dos Estados Unidos) é o maior, com 61% da amostra, estaríamos corretos 61,0% do tempo se designássemos todas as observações a esse grupo. Se escolhemos o critério de chance máxima como o padrão de avaliação, nosso modelo deve ter um desempenho superior a 61% de precisão de classificação para ser aceitável.

Para tentar garantir significância prática, a precisão de classificação alcançada deve exceder o padrão de comparação escolhido em 25%. Assim, devemos selecionar um padrão de comparação, calcular o valor de referência e comparar com a razão de sucesso conseguida.

Todos os níveis de precisão de classificação (razões de sucesso) excedem 85%, o que é consideravelmente maior do que o critério de chance proporcional de 52,4% ou mesmo do critério de chance máxima de 61,0%. Todas as três razões também excedem o valor de referência sugerido desses valores (padrão de comparação mais 25%), que neste caso é de 65,5% (52,4% × 1,25 = 65,5%) para a chance proporcional e 76,3% (61,0% × 1,25 = 76,3%) para a chance máxima. Em todos os casos (amostra de análise, de teste e de validação cruzada), os níveis de precisão de classificação são substancialmente maiores do que os valores de referência, indicando um nível aceitável de precisão de classificação. Além disso, a razão de sucesso para grupos individuais é considerada adequada também.

A medida final de precisão de classificação é o Q de Press, que é uma medida estatística que compara precisão de classificação com um processo aleatório.

A partir da discussão anterior, o cálculo para a amostra de estimação é

$$Q \text{ de Press}_{\text{amostra de estimação}} = \frac{[60-(52\times 2)]^2}{60(2-1)} = 45{,}07$$

E o cálculo para a amostra de validação é

$$Q \text{ de Press}_{\text{amostra de teste}} = \frac{[40-(34\times 2)]^2}{40(2-1)} = 19{,}6$$

Em ambos os casos, os valores calculados excedem o valor crítico de 6,63. Assim, a precisão de classificação para a amostra de análise e, mais importante, para a amostra de validação excede em um nível estatisticamente significante a precisão esperada de classificação por chance.

O pesquisador sempre deve lembrar de tomar cuidado na aplicação de uma amostra de validação com pequenos conjuntos de dados. Nesse caso, a pequena amostra de 40 para validação foi adequada, mas tamanhos maiores são sempre mais desejáveis.

Diagnósticos por casos

Além dos resultados gerais, podemos examinar as observações individuais no que se refere à precisão preditiva e identificar especificamente os casos mal classificados. Nesta operação, podemos encontrar os casos específicos mal classificados para cada grupo nas amostras de análise e de teste e ainda promover uma análise adicional na qual se determine o perfil dos casos mal classificados.

A Tabela 5-11 contém as previsões de grupo para as amostras de análise e de validação e nos permite identificar os casos específicos para cada tipo de má classificação tabulada nas matrizes de classificação (ver Tabela 5-10). Para a amostra de análise, os sete clientes localizados fora dos Estados Unidos que foram mal classificados no grupo de clientes na América do Norte podem ser identificados como os casos 3, 94, 49, 64, 24, 53 e 32. Analogamente, o único cliente dos Estados Unidos que foi mal classificado é identificado como caso 43. Um exame semelhante pode ser feito para a amostra de validação.

Assim que os casos mal classificados são identificados, uma análise adicional pode ser realizada para compreender as razões dessa má classificação. Na Tabela 5-12, os casos mal classificados são combinados a partir das amostras de análise e de validação e então comparados com os casos corretamente classificados. O objetivo é identificar diferenças específicas nas variáveis independentes que possam identificar novas variáveis a serem acrescentadas ou características em comum que devam ser consideradas.

Os cinco casos (tanto na amostra de análise quanto na de validação) mal classificados entre os clientes dos Estados Unidos (grupo 0) têm diferenças significantes em duas das três variáveis independentes na função discriminante (X_{13} e X_{17}), bem como em uma variável não incluída na função discriminante (X_6). Para tal variável, o perfil dos casos mal classificados não é semelhante ao seu grupo correto; logo, não ajuda na classificação. Analogamente, os nove casos mal classificados do grupo 1 (fora dos Estados Unidos) mostram quatro diferenças significantes (X_6, X_{11}, X_{13} e X_{17}), mas apenas X_6 não está na função discriminante. Podemos ver que aqui X_6 funciona contra a precisão de classificação porque os casos mal classificados são mais semelhantes ao grupo incorreto do que ao outro.

(*Continua*)

TABELA 5-11 Previsões de grupo para casos individuais na análise discriminante de dois grupos

Identificação do caso	Grupo real	Escore *Z* discriminante	Grupo previsto	Identificação de caso	Grupo real	Escore *Z* discriminante	Grupo previsto
Amostra de análise							
72	0	–2,10690	0	24	1	–0,60937	0
14	0	–2,03496	0	53	1	–0,45623	0
31	0	–1,98885	0	32	1	–0,36094	0
54	0	–1,98885	0	80	1	–0,14687	1
27	0	–1,76053	0	38	1	–0,04489	1
29	0	–1,76053	0	60	1	–0,04447	1
16	0	–1,71859	0	65	1	0,09785	1
61	0	–1,71859	0	35	1	0,84464	1
79	0	–1,57916	0	1	1	0,98896	1
36	0	–1,57108	0	4	1	1,10834	1
98	0	–1,57108	0	68	1	1,12436	1
58	0	–1,48136	0	44	1	1,34768	1
45	0	–1,33840	0	17	1	1,35578	1
2	0	–1,29645	0	67	1	1,35578	1
52	0	–1,29645	0	33	1	1,42147	1
50	0	–1,24651	0	87	1	1,57544	1
47	0	–1,20903	0	6	1	1,58353	1
88	0	–1,10294	0	46	1	1,60411	1
11	0	–0,74943	0	12	1	1,75931	1
56	0	–0,73978	0	69	1	1,82233	1
95	0	–0,73978	0	86	1	1,82233	1
81	0	–0,72876	0	10	1	1,85847	1
5	0	–0,60845	0	30	1	1,90062	1
37	0	–0,60845	0	15	1	1,91724	1
63	0	–0,38398	0	92	1	1,97960	1
43	0	0,23553	1	7	1	2,09505	1
3	1	–1,65744	0	20	1	2,22839	1
94	1	–1,57916	0	8	1	2,39938	1
49	1	–1,04667	0	100	1	2,62102	1
64	1	–0,67406	0	48	1	2,90178	1
Amostra de teste							
23	0	22,38834	0	25	1	1,47048	1
93	0	–2,03496	0	18	1	1,60411	1
59	0	–1,20903	0	73	1	1,61002	1
85	0	–1,10294	0	21	1	1,69348	1
83	0	–1,03619	0	90	1	1,69715	1
91	0	–0,89292	0	97	1	1,70398	1
82	0	–0,74943	0	40	1	1,75931	1
76	0	–0,72876	0	77	1	1,86055	1
96	0	–0,57335	0	28	1	1,97494	1
13	0	0,13119	1	71	1	2,22839	1
89	0	0,51418	1	19	1	2,28652	1
42	0	0,63440	1	57	1	2,31456	1
78	0	0,63440	1	9	1	2,36823	1
22	1	–2,73303	0	41	1	2,53652	1
74	1	–1,04667	0	26	1	2,59447	1
51	1	0,09785	1	70	1	2,59447	1
62	1	0,94702	1	66	1	2,90178	1
75	1	0,98896	1	34	1	2,97632	1
99	1	1,13130	1	55	1	2,97632	1
84	1	1,30393	1	39	1	3,21116	1

TABELA 5-12 Perfil de observações corretamente classificadas e mal classificadas na análise discriminante de dois grupos

		Escores médios			Teste *t*
Variável dependente: X_4 Região	**Variáveis (Grupo/Perfil)**	**Corretamente classificada**	**Mal classificada**	**Diferença**	**Significância estatística**
EUA/América do Norte		(n = 34)	(n = 5)		
	X_6 Qualidade do produto	8,612	9,340	–0,728	0,000[b]
	X_7 Atividades de comércio eletrônico	3,382	4,380	–0,998	0,068[b]
	X_8 Suporte técnico	5,759	5,280	0,479	0,487
	X_9 Solução de reclamação	5,356	6,140	–0,784	0,149
	X_{10} Anúncio	3,597	4,700	–1,103	0,022
	X_{11} Linha do produto[a]	6,726	6,540	0,186	0,345[b]
	X_{12} Imagem da equipe de venda	4,459	5,460	–1,001	0,018
	X_{13} Preços competitivos[a]	5,609	8,060	–2,451	0,000
	X_{14} Garantia e reclamações	6,215	6,060	0,155	0,677
	X_{15} Novos produtos	5,024	4,420	0,604	0,391
	X_{16} Encomenda e cobrança	4,188	4,540	–0,352	0,329
	X_{17} Flexibilidade de preço[a]	3,568	4,480	–0,912	0,000[b]
	X_{18} Velocidade de entrega	3,826	4,160	–0,334	0,027[b]
Fora da América do Norte		(n = 52)	(n = 9)		
	X_6 Qualidade do produto	6,906	9,156	–2,250	0,000
	X_7 Atividades de comércio eletrônico	3,860	3,289	0,571	0,159[b]
	X_8 Suporte técnico	5,085	5,544	–0,460	0,423
	X_9 Solução de reclamação	5,365	5,822	–0,457	0,322
	X_{10} Anúncio	4,229	3,922	0,307	0,470
	X_{11} Linha do produto[a]	4,954	6,833	–1,879	0,000
	X_{12} Imagem da equipe de venda	5,465	5,467	–1,282E−03	0,998
	X_{13} Preços competitivos[a]	7,960	5,833	2,126	0,000
	X_{14} Garantia e reclamações	5,867	6,400	–0,533	0,007[b]
	X_{15} Novos produtos	5,194	5,778	–0,584	0,291
	X_{16} Encomenda e cobrança	4,267	4,533	–0,266	0,481
	X_{17} Flexibilidade de preço[a]	5,458	3,722	1,735	0,000
	X_{18} Velocidade de entrega	3,881	3,989	–0,108	0,714

Nota: Casos das amostras de análise e validação incluídos para a amostra total de 100.
[a]Variáveis incluídas na função discriminante.
[b]Teste *t* executado com estimativas separadas de variância no lugar de uma estimativa coletiva, pois o teste Levene detectou diferenças significantes nas variações entre os dois grupos.

(*Continuação*)

As descobertas sugerem que os casos mal classificados podem representar um terceiro grupo, pois eles compartilham perfis muito semelhantes nessas variáveis, mais do que acontece nos dois grupos existentes. A administração pode analisar esse grupo quanto a variáveis adicionais ou avaliar se um padrão geográfico entre os casos mal classificados justifica um terceiro grupo.

Pesquisadores devem examinar os padrões em ambos os grupos com o objetivo de entender as características comuns a eles em uma tentativa de definir os motivos para a má classificação.

Estágio 5: Interpretação dos resultados

Após estimar a função, a próxima fase é a interpretação. Este estágio envolve o exame da função para determinar a importância relativa de cada variável independente na discriminação entre os grupos, interpretar a função discriminante com base nas cargas discriminantes, e então fazer o perfil de cada grupo sobre o padrão de valores médios para variáveis identificadas como discriminadoras importantes.

Identificação de variáveis discriminantes importantes

Como anteriormente discutido, cargas discriminantes são consideradas a medida mais adequada de poder discriminante, mas consideraremos também os pesos discriminantes para fins de comparação. Os pesos discriminantes, na forma padronizada ou não, representam a contribuição de cada variável à função discriminante. Contudo, como discutiremos, multicolinearidade entre as variáveis independentes pode causar impacto na interpretação usando somente os pesos.

Cargas discriminantes são calculadas para cada variável independente, mesmo para aquelas que não estão incluídas na função discriminante. Assim, pesos* discriminantes representam o único impacto de cada variável independente e não são restritas apenas ao impacto compartilhado devido à multicolinearidade. Além disso, como elas são relativamente pouco afetadas pela multicolinearidade, elas representam mais precisamente a associação de cada variável com o escore discriminante.

A Tabela 5-13 contém o conjunto inteiro de medidas interpretativas, incluindo pesos discriminantes padronizados e não-padronizados, cargas para a função discriminante, lambda de Wilks e a razão *F* univariada. As 13 variáveis independentes originais foram examinadas pelo procedimento *stepwise*, e três (X_{11}, X_{13} e X_{17}) são suficientemente significantes para serem incluídas na função. Para fins de interpretação, ordenamos as variáveis independentes em termos de suas cargas e valores *F* univariados – ambos indicadores do poder discriminante de cada variável. Sinais dos pesos ou cargas não afetam a ordem; eles simplesmente indicam uma relação positiva ou negativa com a variável dependente.

Análise de lambda de Wilks e o *F* univariado. O lambda de Wilks e o *F* univariado representam os efeitos separados ou univariados de cada variável, não considerando multicolinearidade entre as variáveis independentes. Análogos às correlações bivariadas da regressão múltipla, eles indicam a habilidade de cada variável para discriminar entre os grupos, mas apenas separadamente. Para interpretar qualquer combinação de duas ou mais variáveis independentes, exige-se análise dos pesos ou cargas discriminantes como descrito nas próximas seções.

A Tabela 5-13 mostra que as variáveis (X_{11}, X_{13} e X_{17}) com os três maiores valores *F* (e os menores lambdas de Wilks) eram também as variáveis que entraram na função discriminante. X_6, porém, tinha um efeito discriminante significante quando considerada separadamente, mas tal efeito era compartilhado com as outras três variáveis, de maneira que sozinha ela não contribuía suficientemente para entrar na função discriminante. Todas as demais variáveis tinham valores *F* não-significantes e valores lambda de Wilks correspondentemente elevados.

Análise dos pesos discriminantes. Os pesos discriminantes estão disponíveis em formas não-padronizadas e padronizadas. Os pesos não-padronizados (mais a constante) são usados para calcular o escore discriminante, mas podem ser afetados pela escala da variável independente (exatamente como pesos de regressão múltipla). Assim, os pesos padronizados refletem mais verdadeiramente o impacto de cada variável sobre a função discriminante e são mais apropriados para fins de interpretação. Se for usada estimação simultânea, multicolinearidade entre quaisquer variáveis independentes causará impacto sobre os pesos estimados. No entanto, o impacto da multicolinearidade pode ser até maior para o procedimento *stepwise*, pois ela afeta não somente os pesos mas pode também impedir que uma variável sequer entre na equação.

A Tabela 5-13 fornece os pesos padronizados (coeficientes) para as três variáveis incluídas na função discrimi-

* N. de R. T.: A palavra correta seria "cargas".

TABELA 5-13 Resumo de medidas interpretativas para análise discriminante de dois grupos

	Coeficientes discriminantes		*Cargas discriminantes*		*Lambda de Wilks*	*Razão F univariada*		
Variáveis independentes	**Não padronizados**	**Padronizados**	**Carga**	**Ordenação**	**Valor**	**Valor F**	**Sig.**	**Ordenação**
X_6 Qualidade do produto	NI	NI	–0,418	5	0,801	14,387	0,000	4
X_7 Atividades de comércio eletrônico	NI	NI	0,429	4	0,966	2,054	0,157	6
X_8 Suporte técnico	NI	NI	–0,136	11	0,973	1,598	0,211	7
X_9 Solução de reclamação	NI	NI	–0,181	8	0,986	0,849	0,361	8
X_{10} Anúncio	NI	NI	0,238	7	0,987	0,775	0,382	9
X_{11} Linha do produto	–0,363	–0,417	–0,586	3	0,695	25,500	0,000	3
X_{12} Imagem da equipe de venda	NI	NI	0,164	9	0,856	9,733	0,003	5
X_{13} Preços competitivos	0,398	0,490	0,656	1	0,645	31,992	0,000	1
X_{14} Garantia e reclamações	NI	NI	–0,329	6	0,992	0,453	0,503	11
X_{15} Novos produtos	NI	NI	0,041	13	0,990	0,600	0,442	10
X_{16} Encomenda e cobrança	NI	NI	–0,149	10	0,999	0,087	0,769	13
X_{17} Flexibilidade de preço	0,749	0,664	0,653	2	0,647	31,699	0,000	2
X_{18} Velocidade de entrega	NI	NI	–0,060	12	0,997	0,152	0,698	12

NI = Não incluído na função discriminante estimada

nante. O impacto da multicolinearidade sobre os pesos pode ser visto ao se examinar X_{13} e X_{17}. Essas duas variáveis têm poder discriminante essencialmente equivalente quando vistas nos testes lambda de Wilks e F univariado. Seus pesos discriminantes, contudo, refletem um impacto sensivelmente maior para X_{17} do que para X_{13}, que agora é mais comparável com X_{11}. Essa mudança em importância relativa é devida à multicolinearidade entre X_{13} e X_{11}, o que reduz o efeito único de X_{13} e assim diminui os pesos discriminantes também.

Interpretação da função discriminante com base nas cargas discriminantes

As cargas discriminantes, em contraste com os pesos discriminantes, são menos afetadas pela multicolinearidade e, portanto, mais úteis para a interpretação. Além disso, como cargas são calculadas para todas as variáveis, elas fornecem uma medida interpretativa até mesmo para variáveis não incluídas na função discriminante. Uma regra prática anterior indicava que cargas acima de ± 0,40 deveriam ser usadas para identificar variáveis discriminantes importantes.

As cargas das três variáveis da função discriminante (ver Tabela 5-13) são as três maiores, e todas excedem ± 0,40, garantindo assim inclusão no processo de interpretação. Duas variáveis adicionais (X_6 e X_7), porém, também têm cargas acima da referência ± 0,40. A inclusão de X_6 não é inesperada, como era a quarta variável com efeito discriminante univariado, mas não foi incluída na função discriminante devido à multicolinearidade (como mostrado no Capítulo 3, Análise Fatorial, onde X_6 e X_{13} formavam um fator). X_7, porém, apresenta outra situação; ela não tinha um efeito univariado significante. A combinação das três variáveis na função discriminante criou um efeito que é associado com X_7, mas X_7 não acrescenta qualquer poder discriminante adicional. Com relação a isso, X_7 é descritiva da função discriminante mesmo não sendo incluída nem tendo um efeito univariado significante.

Interpretar a função discriminante e sua discriminação entre esses dois grupos exige que o pesquisador considere todas essas cinco variáveis. Na medida em que elas caracterizam ou descrevem a função discriminante, todas representam algum componente da mesma.

Com as variáveis discriminantes identificadas e a função discriminante descrita em termos daquelas variáveis com cargas suficientemente elevadas, o pesquisador prossegue então para o perfil de cada grupo sobre essas variáveis para compreender as diferenças entre as mesmas.

Os três efeitos mais fortes na função discriminante, que são geralmente comparáveis com base nos valores de carga, são X_{13} (Preços competitivos), X_{17} (Flexibilidade de preço) e X_{11} (Linha do produto). X_7 (Atividades de comércio eletrônico) e o efeito de X_6 (Qualidade do produto) podem ser adicionados aos efeitos de X_{13}. Obviamente, diversos fatores diferentes estão sendo combinados para diferenciar entre os grupos, exigindo assim mais definição de perfil dos grupos para se entenderem as diferenças.

Perfil das variáveis discriminantes

O pesquisador está interessado em interpretações das variáveis individuais que têm significância estatística e prática. Tais interpretações são conseguidas primeiramente identificando-se as variáveis com substantivo poder discriminatório (ver a discussão anterior) e em seguida entendendo-se o que o grupo distinto diz cada variável indicada.

Como descrito no Capítulo 1, escores maiores nas variáveis independentes indicam percepções mais favoráveis da HBAT sobre aquele atributo (exceto para X_{13}, onde escores menores são preferíveis). Retornando à Tabela 5-5, vemos diversos perfis entre os dois grupos sobre essas cinco variáveis.

- O grupo 0 (clientes nos Estados Unidos/América do Norte) têm percepções maiores sobre três variáveis: X_6 (Qualidade do produto), X_{13}* (Preços competitivos) e X_{11} (Linha do produto).
- O grupo 1 (clientes fora da América do Norte) têm percepções maiores nas outras duas variáveis: X_7 (Atividades de comércio eletrônico) e X_{17} (Flexibilidade de preço).

Olhando esses dois perfis, podemos perceber que os clientes dos EUA/América do Norte têm percepções muito melhores dos produtos HBAT, enquanto os demais clientes se sentem melhor com questões sobre preço e comércio eletrônico. Note que X_6 e X_{13}, ambas com percepções mais elevadas entre os clientes dos EUA/América do Norte, formam o fator Valor do produto desenvolvido no Capítulo 3. A administração deveria usar esses resultados para desenvolver estratégias que acentuem esses pontos fortes e desenvolver outras vantagens para fins de complementação.

O perfil médio também ilustra a interpretação dos sinais (positivos e negativos) nos pesos e as cargas discriminantes. Os sinais refletem o perfil médio relativo dos dois grupos. Os sinais positivos, neste exemplo, são associados com variáveis que têm escores maiores para o grupo 1. Os pesos e cargas negativas são para aquelas variáveis com o padrão oposto (i.e., valores maiores no grupo 0). Logo, os sinais indicam o padrão entre os grupos.

* N. de R. T.: A tabela indica o contrário, ou seja, a média de X_{13} é maior no grupo 1 (7,418 versus 5,600).

Estágio 6: Validação dos resultados

O estágio final aborda a validade interna e externa da função discriminante. O principal meio de validação é pelo uso da amostra de validação e a avaliação de sua precisão preditiva. Desse modo, a validade é estabelecida se a função discriminante classifica, em um nível aceitável, observações que não foram usadas no processo de estimação. Se a amostra de validação é obtida a partir da amostra original, então essa abordagem estabelece validade interna. Se uma outra amostra separada, talvez de uma outra população ou de outro segmento da população, forma a amostra de validação, então isso corresponde a uma validação externa dos resultados discriminantes.

Em nosso exemplo, a amostra de teste surge a partir da amostra original. Como anteriormente discutido, a precisão de classificação (razões de sucesso) para as amostras de teste e de validação cruzada estava muito acima das referências em todas as medidas de precisão preditiva. Como tal, a análise estabelece validade interna. Para o propósito de validade externa, amostras adicionais devem ser extraídas de populações relevantes e a precisão de classificação deve ser avaliada em tantas situações quanto possível.

O pesquisador é encorajado a estender o processo de validação por meio de perfis expandidos dos grupos e o possível uso de amostras adicionais para estabelecer a validade externa. Idéias adicionais da análise de casos mal classificados podem sugerir variáveis extras que podem melhorar ainda mais o modelo discriminante.

Uma visão gerencial

A análise discriminante de clientes HBAT, baseada em localização geográfica (dentro ou fora da América do Norte), identificou um conjunto de diferenças em percepção que pode fornecer uma distinção mais sucinta e poderosa entre os dois grupos. Várias descobertas importantes incluem as seguintes:

- Diferenças são encontradas em um subconjunto de apenas cinco percepções, o que permite uma concentração sobre as variáveis-chave, não tendo que se lidar com o conjunto inteiro. As variáveis identificadas como discriminantes entre os grupos (listadas em ordem de importância) são X_{13} (Preços competitivos), X_{17} (Flexibilidade de preço), X_{11} (Linha do produto), X_7 (Atividades de comércio eletrônico) e X_6 (Qualidade do produto).
- Os resultados também indicam que as empresas localizadas nos Estados Unidos têm melhores percepções da HBAT do que suas contrapartes internacionais em termos de valor e linha de produto, enquanto os clientes que não são norte-americanos têm uma percepção mais favorável sobre flexibilidade de preço e atividades de comércio eletrônico. Essas percepções podem resultar de uma maior similaridade entre compradores norte-americanos, enquanto clientes internacionais acham a política de preços em sintonia com suas necessidades.
- Os resultados, que são altamente significantes, fornecem ao pesquisador a habilidade de identificar corretamente a estratégia de compra usada, com base nessas percepções, 85% do tempo. Esse elevado grau de consistência gera confiança no desenvolvimento de estratégias baseadas em tais resultados.
- A análise das empresas mal classificadas revelou um pequeno número de empresas que pareciam "deslocadas". Identificar tais empresas pode identificar associações não tratadas por localização geográfica (p.ex., mercados no lugar de apenas localização física) ou outras características de firmas ou de mercado que são associadas com localização geográfica.

Portanto, conhecer a localização de uma firma dá idéias-chave sobre suas percepções da HBAT e, mais importante, como os dois grupos de clientes diferem, de forma que a administração pode empregar uma estratégia para acentuar as percepções positivas em suas negociações com esses clientes e assim solidificar sua posição.

UM EXEMPLO ILUSTRATIVO DE TRÊS GRUPOS

Para ilustrar a aplicação de uma análise discriminante de três grupos, novamente usamos a base de dados HBAT. No exemplo anterior, estávamos interessados na discriminação entre apenas dois grupos, de modo que conseguimos desenvolver uma única função discriminante e um escore de corte para dividir os dois grupos. No exemplo de três grupos, é necessário desenvolver duas funções discriminantes separadas para distinguir entre os três grupos. A primeira função separa um grupo dos outros dois, e a segunda separa os dois grupos restantes. Como no exemplo anterior, os seis estágios do processo de construção do modelo são discutidos.

Estágio 1: Objetivos da análise discriminante

O objetivo da HBAT nessa pesquisa é determinar a relação entre as percepções que as empresas têm da HBAT e o período de tempo em que uma empresa é cliente de HBAT.

Um dos paradigmas emergentes em marketing é o conceito de uma relação com cliente, baseada no estabelecimento de uma mútua parceria entre empresas ao longo de repetidas transações. O processo de desenvolvimento de uma relação implica a formação de metas e valores compartilhados, que devem coincidir com percepções melhoradas de HBAT. Portanto, a formação bem-sucedida de uma relação deve ser entendida por meio de per-

(*Continua*)

(*Continuação*)
cepções melhores de HBAT ao longo do tempo. Nessa análise, as firmas são agrupadas conforme sua situação como clientes HBAT. Se HBAT foi bem-sucedida no estabelecimento de relações com seus clientes, então as percepções sobre a HBAT irão melhorar em cada situação como cliente HBAT.

Estágio 2: Projeto de pesquisa para análise discriminante

Para testar essa relação, uma análise discriminante é executada para estabelecer se existem diferenças em percepções entre grupos de clientes com base na extensão da relação de clientela. Se for o caso, a HBAT estará então interessada em ver se diferentes perfis justificam a proposição de que a HBAT teve sucesso no melhoramento de percepções entre clientes estabelecidos, um passo necessário na formação de relações com a clientela.

Seleção de variáveis dependente e independentes

Além das variáveis dependentes não-métricas (categóricas) definindo grupos de interesse, a análise discriminante também requer um conjunto de variáveis independentes métricas que são consideradas fornecedoras de base para discriminação ou diferenciação entre os grupos.

Uma análise discriminante de três grupos é realizada usando X_1 (Tipo de cliente) como a variável dependente e as percepções de HBAT por parte dessas firmas (X_6 a X_{18}) como as variáveis independentes. Note que X_1 difere da variável dependente no exemplo de dois grupos no sentido de que ela tem três categorias nas quais classificar o tempo de permanência como cliente de HBAT (1 = menos que 1 ano, 2 = 1 a 5 anos, e 3 = mais de 5 anos).

Tamanho amostral e divisão da amostra

Questões relativas ao tamanho da amostra são particularmente importantes com análise discriminante devido ao foco não apenas no tamanho geral da amostra, mas também no tamanho amostral por grupo. Juntamente com a necessidade de uma divisão da amostra para obter uma amostra de validação, o pesquisador deve considerar cuidadosamente o impacto da divisão amostral em termos do tamanho geral e do tamanho de cada um dos grupos.

A base de dados da HBAT tem uma amostra de 100, a qual será novamente particionada em amostras de análise e de validação de 60 e 40 casos, respectivamente. Na amostra de análise, a proporção de casos por variáveis independentes é quase 5:1, o limite inferior recomendado. Mais importante, na amostra de análise, apenas um grupo, com 13 observações, fica abaixo do nível recomendado de 20 casos por grupo. Apesar de o tamanho do grupo exceder 20 se a amostra inteira for usada na fase de análise, a necessidade de validação dita a criação da amostra de teste. Os três grupos são de tamanhos relativamente iguais (22, 13 e 25), evitando assim qualquer necessidade de igualar os tamanhos dos grupos. A análise procede com atenção para a classificação e interpretação desse pequeno grupo de 13 observações.

Estágio 3: Suposições da análise discriminante

Como no caso do exemplo de dois grupos, as suposições de normalidade, linearidade e colinearidade das variáveis independentes já foram discutidas detalhadamente no Capítulo 2. A análise feita no Capítulo 2 indicou que as variáveis independentes atendem essas suposições em níveis adequados para viabilizar a continuidade da análise sem ações corretivas adicionais. A suposição remanescente, a igualdade de matrizes de variância/covariância ou de dispersão, também é abordada no Capítulo 2.

O teste M de Box avalia a similaridade das matrizes de dispersão das variáveis independentes entre os três grupos (categorias). O teste estatístico indicou diferenças no nível de significância de 0,09. Neste caso, as diferenças entre grupos são não-significantes e nenhuma ação corretiva se faz necessária. Além disso, não se espera qualquer impacto sobre os processos de estimação e classificação.

Estágio 4: Estimação do modelo discriminante e avaliação do ajuste geral

Como no exemplo anterior, começamos nossa análise revisando as médias de grupo e os desvios-padrão para ver se os grupos são significantemente diferentes em alguma variável. Com essas diferenças em mente, empregamos em seguida um processo de estimação *stepwise* para obter as funções discriminantes e completamos o processo avaliando precisão de classificação com diagnósticos gerais e por casos.

Avaliação de diferenças de grupos

Identificar as variáveis mais discriminantes com três ou mais grupos é mais problemático do que na situação com dois grupos. Para três ou mais grupos, as medidas típicas de significância para diferenças em grupos (ou seja, lambda de Wilks e o teste *F*) avaliam apenas as diferenças gerais e não garantem que cada grupo é significante em relação aos demais. Assim, quando examinar variáveis quanto a suas diferenças gerais entre os grupos, certifique-se também de tratar das diferenças individuais de grupos.

A Tabela 5-14 dá as médias de grupos, lambda de Wilks, razões *F* univariadas (ANOVAs simples) e D^2 mínimo

(*Continua*)

(*Continuação*)

de Mahalanobis para cada variável independente. A revisão dessas medidas revela o seguinte:

- Sobre uma base univariada, aproximadamente metade (7 entre 13) das variáveis exibe diferenças significantes entre as médias dos grupos. As variáveis com diferenças significantes incluem X_6, X_9, X_{11}, X_{13}, X_{16}, X_{17} e X_{18}.
- Apesar de maior significância estatística corresponder a uma maior discriminação geral (ou seja, as variáveis mais significantes têm os menores lambdas de Wilks), ela nem sempre corresponde à maior discriminação entre todos os grupos.
  - A inspeção visual das médias dos grupos revela que quatro das variáveis com diferenças significantes (X_{13}, X_{16}, X_{17} e X_{18}) diferenciam apenas um grupo versus os outros dois grupos [p.ex., X_{18} tem diferenças significantes somente nas médias entre o grupo 1 (3,059) versus grupos 2 (4,246) e 3 (4,288)]. Essas variáveis têm um papel limitado em análise discriminante por fornecerem discriminação apenas em um subconjunto de grupos.
  - Três variáveis (X_6, X_9 e X_{11}) fornecem alguma discriminação, em vários graus, entre todos os grupos simultaneamente. Uma ou mais dessas variáveis podem ser usadas em combinação com as quatro variáveis precedentes para criar uma variável estatística com discriminação máxima.
- O valor D^2 de Mahalanobis fornece uma medida do grau de discriminação entre grupos. Para cada variável, o D^2 mínimo de Mahalanobis é a distância entre os dois grupos mais próximos. Por exemplo, X_{11} tem o maior valor D^2 e é a variável com as maiores diferenças entre todos os três grupos. Analogamente, X_{18}, uma variável com pequenas diferenças entre dois dos grupos, tem um pequeno valor D^2. Com três ou mais grupos, o D^2 mínimo de Mahalanobis é importante na identificação da variável que dá a maior diferença entre os dois grupos mais parecidos.

Todas essas medidas se combinam para ajudar a identificar os conjuntos de variáveis que formam as funções discriminantes, como descritos na próxima seção. Quando mais de uma função é criada, cada uma fornece discriminação entre conjuntos de grupos. No exemplo simples do início deste capítulo, uma variável discriminou entre os grupos 1 versus 2 e 3, sendo que a outra discriminou entre os grupos 2 versus 3 e 1. Esse é um dos principais benefícios que surgem do uso da análise discriminante.

Estimação da função discriminante

O procedimento *stepwise* é realizado da mesma maneira do exemplo de dois grupos, com todas as variáveis inicialmente excluídas do modelo. O procedimento então seleciona a variável que tem uma diferença estatisticamente significante nos grupos enquanto maximiza a distância de Mahalanobis (D^2) entre os dois grupos mais próximos. Desta maneira, variáveis estatisticamente significantes são selecionadas de modo a maximizarem a discriminação entre os grupos mais semelhantes em cada estágio.

Este processo continua enquanto variáveis adicionais fornecerem discriminação estatisticamente significante além daquelas diferenças já explicadas pelas variáveis na função discriminante. Uma variável pode ser removida se alta multicolinearidade com variáveis independentes na função discriminante faz com que sua significância caia abaixo do nível para remoção (0,10).

TABELA 5-14 Estatísticas descritivas de grupos e testes de igualdade para a amostra de estimação na análise discriminante de três grupos

	Médias de grupo da variável dependente: X_1 Tipo de cliente			***Teste de igualdade de médias de grupos[a]***			***D^2 mínimo de Mahalanobis***	
Variáveis independentes	**Grupo 1: Menos que 1 ano (n = 22)**	**Grupo 2: 1 a 5 anos (n = 13)**	**Grupo 3: Mais de 5 anos (n = 25)**	**Lambda de Wilks**	**Valor F**	**Significância**	**D^2 mínimo**	**Entre grupos**
X_6 Qualidade do produto	7,118	6,785	9,000	0,469	32,311	0,000	0,121	1 e 2
X_7 Atividades de comércio eletrônico	3,514	3,754	3,412	0,959	1,221	0,303	0,025	1 e 3
X_8 Suporte técnico	4,959	5,615	5,376	0,973	0,782	0,462	0,023	2 e 3
X_9 Solução de reclamação	4,064	5,900	6,300	0,414	40,292	0,000	0,205	2 e 3
X_{10} Anúncio	3,745	4,277	3,768	0,961	1,147	0,325	0,000	1 e 3
X_{11} Linha do produto	4,855	5,577	7,056	0,467	32,583	0,000	0,579	1 e 2
X_{12} Imagem da equipe de venda	4,673	5,346	4,836	0,943	1,708	0,190	0,024	1 e 3
X_{13} Preços competitivos	7,345	7,123	5,744	0,751	9,432	0,000	0,027	1 e 2
X_{14} Garantia e reclamações	5,705	6,246	6,072	0,916	2,619	0,082	0,057	2 e 3
X_{15} Novos produtos	4,986	5,092	5,292	0,992	0,216	0,807	0,004	1 e 2
X_{16} Encomenda e cobrança	3,291	4,715	4,700	0,532	25,048	0,000	0,000	2 e 3
X_{17} Flexibilidade de preço	4,018	5,508	4,084	0,694	12,551	0,000	0,005	1 e 3
X_{18} Velocidade de entrega	3,059	4,246	4,288	0,415	40,176	0,000	0,007	2 e 3

[a]Lambda de Wilks (estatística *U*) e razão *F* univariada com 2 e 57 graus de liberdade.

Estimação *stepwise*: adição da primeira variável, X_{11}. Os dados na Tabela 5-14 mostram que a primeira variável a entrar no modelo é X_{11} (Linha do produto), pois ela atende aos critérios para diferenças estatisticamente significantes nos grupos e tem o maior valor D^2 (o que significa que ela tem a maior separação entre os dois grupos mais parecidos).

Os resultados de adicionar X_{11} como a primeira variável no processo *stepwise* são mostrados na Tabela 5-15. O ajuste geral do modelo é significante e todos os grupos são significantemente distintos, apesar de os grupos 1 (menos de um ano) e 2 (de um a cinco anos) terem a menor diferença entre eles (ver seção abaixo detalhando as diferenças de grupos).

Com a menor diferença entre os grupos 1 e 2, o procedimento discriminante selecionará agora uma variável que maximiza aquela diferença enquanto pelo menos mantém as demais. Se voltarmos à Tabela 5-14, perceberemos que quatro variáveis (X_9, X_{16}, X_{17} e X_{18}) tinham diferenças significantes, com substanciais distinções entre os grupos 1 e 2. Olhando a Tabela 5-15, vemos que essas quatro variáveis têm o maior valor D^2 mínimo, e em cada caso é para a diferença entre os grupos 2 e 3 (o que significa que os grupos 1 e 2 não são os mais pa-

(*Continua*)

TABELA 5-15 Resultados do passo 1 da análise discriminante *stepwise* de três grupos

Ajuste geral do modelo

	Valor	Valor *F*	Graus de liberdade	Significância
Lambda de Wilks	0,467	32,583	2,57	0,000

Variável adicionada/removida no passo 1

		F		
Variável adicionada	D^2 mínimo	Valor	Significância	Entre grupos
X_{11} Linha de produto	0,579	4,729	0,000	Menos de 1 ano e de 1 a 5 anos

Nota: Em cada passo, a variável que maximiza a distância Mahalanobis entre os dois grupos mais próximos é adicionada.

Variáveis na análise após o passo 1

Variável	Tolerância	*F* para remover	D^2	Entre grupos
X_{11} Linha de produto	1,000	32,583	NA	NA

NA = Não aplicável

Variáveis fora da análise após o passo 1

Variável	Tolerância	Tolerância mínima	F para entrar	D^2 mínimo	Entre grupos
X_6 Qualidade de produto	1,000	1,000	17,426	0,698	Menos de 1 ano e de 1 a 5 anos
X_7 Atividades de comércio eletrônico	0,950	0,950	1,171	0,892	Menos de 1 ano e de 1 a 5 anos
X_8 Suporte técnico	0,959	0,959	0,733	0,649	Menos de 1 ano e de 1 a 5 anos
X_9 Solução de reclamação	0,847	0,847	15,446	2,455	De 1 a 5 anos e mais de 5 anos
X_{10} Anúncio	0,998	0,998	1,113	0,850	Menos de 1 ano e de 1 a 5 anos
X_{12} Imagem da equipe de venda	0,932	0,932	3,076	1,328	Menos de 1 ano e de 1 a 5 anos
X_{13} Preços competitivos	0,882	0,882	2,299	0,839	Menos de 1 ano e de 1 a 5 anos
X_{14} Garantia e reclamações	0,849	0,849	0,647	0,599	Menos de 1 ano e de 1 a 5 anos
X_{15} Novos produtos	0,993	0,993	0,415	0,596	Menos de 1 ano e de 1 a 5 anos
X_{16} Encomenda e cobrança	0,943	0,943	12,176	2,590	De 1 a 5 anos e mais de 5 anos
X_{17} Flexibilidade de preço	0,807	0,807	17,300	3,322	De 1 a 5 anos e mais de 5 anos
X_{18} Velocidade de entrega	0,773	0,773	19,020	2,988	De 1 a 5 anos e mais de 5 anos

Teste de significância de diferenças de grupos após o passo 1[a]

X_1 Tipo de cliente		Menos de 1 ano	De 1 a 5 anos
De 1 a 5 anos	*F*	4,729	
	Sig.	0,034	
Mais de 5 anos	*F*	62,893	20,749
	Sig.	0,000	0,000

[a]1 e 57 graus de liberdade.

(*Continuação*)

recidos depois de acrescentar aquela variável). Assim, adicionar qualquer uma dessas variáveis afeta muito as diferenças entre os grupos 1 e 2, o par que era mais parecido depois que X_{11} foi adicionada no primeiro passo. O procedimento escolherá X_{17} porque ela criará a maior distância entre os grupos 2 e 3.

Estimação *stepwise*: Adição da segunda variável, X_{17}. A Tabela 5-16 detalha o segundo passo do procedimento *stepwise*: o acréscimo de X_{17} (Flexibilidade de preço) à função discriminante. A discriminação entre grupos aumentou, como refletido em um menor valor lambda de Wilks e no aumento do D^2 mínimo (de 0,467 para 0,288). As diferenças de grupos, geral e individuais, ainda são estatisticamente significantes. O acréscimo de X_{17} aumentou as distinções entre os grupos 1 e 2 consideravelmente, de forma que agora os dois grupos mais parecidos são 2 e 3.

Das variáveis fora da equação, apenas X_6 (Qualidade de produto) satisfaz o nível de significância necessário

(*Continua*)

TABELA 5-16 Resultados do passo 2 da análise discriminante *stepwise* de três grupos

Ajuste geral do modelo

	Valor	Valor *F*	Graus de liberdade	Significância
Lambda de Wilks	0,288	24,139	4, 112	0,000

Variável adicionada/removida no passo 2

		F		
Variável adicionada	D^2 mínimo	Valor	Significância	Entre grupos
X_{17} Flexibilidade de preço	3,322	13,958	0,000	De 1 a 5 anos e mais de 5 anos

Nota: Em cada passo, a variável que maximiza a distância Mahalanobis entre os dois grupos mais próximos é adicionada.

Variáveis na análise após o passo 2

Variável	Tolerância	*F* para remover	D^2	Entre grupos
X_{11} Linha de produto	0,807	39,405	0,005	Menos de 1 ano e mais de 5 anos
X_{17} Flexibilidade de preço	0,807	17,300	0,579	Menos de 1 ano e de 1 a 5 anos

Variáveis fora da análise após o passo 2

Variável	Tolerância	Tolerância mínima	*F* para entrar	D^2 mínimo	Entre grupos
X_6 Qualidade de produto	0,730	0,589	24,444	6,071	Menos de 1 ano e de 1 a 5 anos
X_7 Atividades de comércio eletrônico	0,880	0,747	0,014	3,327	Menos de 1 ano e de 1 a 5 anos
X_8 Suporte técnico	0,949	0,791	1,023	3,655	Menos de 1 ano e de 1 a 5 anos
X_9 Solução de reclamação	0,520	0,475	3,932	3,608	Menos de 1 ano e de 1 a 5 anos
X_{10} Anúncio	0,935	0,756	0,102	3,348	Menos de 1 ano e de 1 a 5 anos
X_{12} Imagem da equipe de venda	0,884	0,765	0,662	3,342	Menos de 1 ano e de 1 a 5 anos
X_{13} Preços competitivos	0,794	0,750	0,989	3,372	Menos de 1 ano e de 1 a 5 anos
X_{14} Garantia e reclamações	0,868	0,750	2,733	4,225	Menos de 1 ano e de 1 a 5 anos
X_{15} Novos produtos	0,963	0,782	0,504	3,505	Menos de 1 ano e de 1 a 5 anos
X_{16} Encomenda e cobrança	0,754	0,645	2,456	3,323	Menos de 1 ano e de 1 a 5 anos
X_{18} Velocidade de entrega	0,067	0,067	3,255	3,598	Menos de 1 ano e de 1 a 5 anos

Teste de significância de diferenças de grupos após o passo 2[a]

X_1 Tipo de cliente		Menos de 1 ano	De 1 a 5 anos
De 1 a 5 anos	*F*	21,054	
	Sig.	0,000	
Mais de 5 anos	*F*	39,360	13,958
	Sig.	0,000	0,000

[a]2 e 56 graus de liberdade.

(*Continuação*)
para consideração. Se acrescentada, o D^2 mínimo será agora entre os grupos 1 e 2.

Estimação *stepwise*: Adição das terceira e quarta variáveis, X_6 e X_{18}. Como anteriormente observado, X_6 se torna a terceira variável adicionada à função discriminante. Depois que X_6 foi acrescentada, apenas X_{18} exibe uma significância estatística nos grupos (*Nota*: Os detalhes sobre o acréscimo de X_6 no terceiro passo não são mostrados por questão de espaço).

A variável final adicionada no passo 4 é X_{18} (ver Tabela 5-17), com a função discriminante incluindo agora quatro variáveis (X_{11}, X_{17}, X_6 e X_{18}). O modelo geral é significante, com o lambda de Wilks diminuindo para 0,127. Além disso, existem diferenças significantes entre todos os grupos individuais.

Com essas quatro variáveis na função discriminante, nenhuma outra variável exibe a significância estatística necessária para inclusão, e o processo *stepwise* está

(*Continua*)

TABELA 5-17 Resultados do passo 4 da análise discriminante *stepwise* de três grupos

Ajuste geral do modelo

	Valor	Valor *F*	Graus de liberdade	Significância
Lambda de Wilks	0,127	24,340	8, 108	0,000

Variável adicionada/removida no passo 4

		F		
Variável adicionada	D^2 mínimo	Valor	Significância	Entre grupos
X_{18} Velocidade de entrega	6,920	13,393	0,000	Menos de 1 ano e de 1 a 5 anos

Nota: Em cada passo, a variável que maximiza a distância Mahalanobis entre os dois grupos mais próximos é adicionada.

Variáveis na análise após o passo 4

Variável	Tolerância	*F* para remover	D^2	Entre grupos
X_{11} Linha de produto	0,075	0,918	6,830	Menos de 1 ano e de 1 a 5 anos
X_{17} Flexibilidade de preço	0,070	1,735	6,916	Menos de 1 ano e de 1 a 5 anos
X_6 Qualidade do produto	0,680	27,701	3,598	De 1 a 5 anos e mais de 5 anos
X_{18} Velocidade de entrega	0,063	5,387	6,071	Menos de 1 ano e de 1 a 5 anos

Variáveis fora da análise após o passo 4

Variável	Tolerância	Tolerância mínima	*F* para entrar	D^2 mínimo	Entre grupos
X_7 Atividades de comércio eletrônico	0,870	0,063	0,226	6,931	Menos de 1 ano e de 1 a 5 anos
X_8 Suporte técnico	0,940	0,063	0,793	7,164	Menos de 1 ano e de 1 a 5 anos
X_9 Solução de reclamação	0,453	0,058	0,292	7,019	Menos de 1 ano e de 1 a 5 anos
X_{10} Anúncio	0,932	0,063	0,006	6,921	Menos de 1 ano e de 1 a 5 anos
X_{12} Imagem da equipe de venda	0,843	0,061	0,315	7,031	Menos de 1 ano e de 1 a 5 anos
X_{13} Preços competitivos	0,790	0,063	0,924	7,193	Menos de 1 ano e de 1 a 5 anos
X_{14} Garantia e reclamações	0,843	0,063	2,023	7,696	Menos de 1 ano e de 1 a 5 anos
X_{15} Novos produtos	0,927	0,062	0,227	7,028	Menos de 1 ano e de 1 a 5 anos
X_{16} Encomenda e cobrança	0,671	0,062	1,478	7,210	Menos de 1 ano e de 1 a 5 anos

Teste de significância de diferenças de grupos após o passo 4[a]

X_1 Tipo de cliente		Menos de 1 ano	De 1 a 5 anos
De 1 a 5 anos	*F*	13,393	
	Sig.	0,000	
Mais de 5 anos	*F*	56,164	18,477
	Sig.	0,000	0,000

[a]4 e 54 graus de liberdade.

(*Continuação*)
concluído em termos de acréscimo de variáveis. Porém, o procedimento inclui também um exame da significância de cada variável para que a mesma seja mantida na função discriminante. Neste caso, o "*F* para remover" para X_{11} e X_{17} é não-significante (0,918 e 1,735, respectivamente), indicando que uma ou ambas são candidatas para remoção da função discriminante.

Estimação *stepwise*: Remoção de X_{17} e X_{11}. Quando X_{18} é adicionada ao modelo no quarto passo (ver a discussão anterior), X_{11} tinha o menor valor "*F* para remover" (0,918), fazendo com que o procedimento *stepwise* eliminasse aquela variável da função discriminante no quinto passo (detalhes sobre este passo 5 são omitidos por questões de espaço). Agora com três variáveis na função discriminante (X_{11}, X_6 e X_{18}), o ajuste geral do modelo ainda é significante e o lambda de Wilks aumentou só um pouco para 0,135. Todos os grupos são significantemente diferentes. Nenhuma variável atinge o nível necessário de significância estatística para ser adicionada à função discriminante, e mais uma variável (X_{11}*) tem um valor "*F* para remover" de 2,552, o que indica que ela também pode ser eliminada da função.

A Tabela 5-18 contém os detalhes do passo 6 do procedimento *stepwise*, onde X_{11} também é removida da função discriminante, restando apenas X_6 e X_{18}. Mesmo com a remoção da segunda variável (X_{11}), o modelo geral ainda é significante e o lambda de Wilks é consideravelmente pequeno (0,148). Devemos observar que este modelo de duas variáveis, X_6 e X_{18}, é um melhoramento em relação ao primeiro modelo de duas variáveis, X_{11} e X_{17}, formado no passo 2 (lambda de Wilks é 0,148 contra o valor do primeiro modelo de 0,288 e todas as diferenças individuais de grupos são muito maiores). Sem variáveis alcançando o nível necessário de significância para adição ou remoção, o procedimento *stepwise* é encerrado.

Resumo do processo de estimação *stepwise*. As funções discriminantes estimadas são composições lineares semelhantes a uma reta de regressão (ou seja, elas são uma combinação linear de variáveis). Assim como uma reta de regressão é uma tentativa de explicar a máxima variação em sua variável dependente, essas composições lineares tentam explicar as variações ou diferenças na variável categórica dependente. A primeira função discriminante é desenvolvida para explicar a maior variação (diferença) nos grupos discriminantes. A segunda função discriminante, que é ortogonal e independente da primeira, explica o maior percentual da variância remanescente (residual) depois que a variância para a primeira função é removida.

A informação fornecida na Tabela 5-19 resume os passos da análise discriminante de três grupos, com os seguintes resultados:

- X_6 e X_{18} são as duas variáveis na função discriminante final, apesar de X_{11} e X_{17} terem sido acrescentadas nos dois primeiros passos e então removidas depois que X_6 e X_{18} foram adicionadas. Os coeficientes não-padronizados e padronizados (pesos) da função discriminante e a matriz estrutural das cargas discriminantes, rotacionadas e não-rotacionadas, também foram fornecidos. A rotação das cargas discriminantes facilita a interpretação da mesma maneira que fatores foram simplificados para interpretação via rotação (ver Capítulo 3 para uma discussão mais detalhada sobre rotação). Examinamos em pormenores as cargas rotacionadas e não-rotacionadas no passo 5.
- A discriminação aumentou com a adição de cada variável (como evidenciado pela diminuição no lambda de Wilks), mesmo com apenas duas variáveis restando no modelo final. Comparando o lambda de Wilks final para a análise discriminante (0,148) com o lambda de Wilks (0,414**) para o melhor resultado de uma única variável, X_9**, vemos que uma melhora acentuada é obtida ao se usar exatamente duas variáveis nas funções discriminantes no lugar de uma única variável.
- A qualidade de ajuste geral para o modelo discriminante é estatisticamente significante e ambas as funções são estatisticamente significantes também. A primeira função explica 91,5% da variância explicada pelas duas funções, com a variância remanescente (8,5%) devida à segunda função. A variância total explicada pela primeira função é $0{,}893^2$, ou 79,7%. A próxima função explica $0{,}517^2$ ou 26,7% da variância remanescente (20,3%). Portanto, a variância total explicada por ambas as funções é de 85,1% [79,7% + (26,7% × 0,203)] da variação total na variável dependente.

Ainda que ambas as funções sejam estatisticamente significantes, o pesquisador sempre deve garantir que as funções discriminantes forneçam diferenças entre todos os grupos. É possível ter funções estatisticamente significantes, mas ter pelo menos um par de grupos que não sejam estatisticamente distintos (i.e., não discriminados entre eles). Este problema se torna especialmente predominante quando o número de grupos aumenta ou vários grupos pequenos são incluídos na análise.

A última seção da Tabela 5-18 fornece os testes de significância para diferenças de grupos entre cada par de grupos (p.ex., grupo 1 versus grupo 2, grupo 1 versus grupo 3 etc.). Todos os pares de grupos mostraram diferenças estatisticamente significantes, denotando que as funções discriminantes criaram separação não apenas em um sentido geral, mas também para cada grupo também. Examinamos os centróides de grupos graficamente em uma seção posterior.

* N. de R. T.: Provavelmente trata-se de X_{17}, uma vez que X_{11} já fora removida.

* N. de R. T.: Na verdade, seria X_{11} com lambda de Wilks igual a 0,467.

TABELA 5-18 Resultados do passo 6 da análise discriminante *stepwise* de três grupos

Ajuste geral do modelo

	Valor	Valor *F*	Graus de liberdade	Significância
Lambda de Wilks	0,148	44,774	4, 112	0,000

Variável adicionada/removida no passo 6

		F		
Variável removida	**D^2 mínimo**	**Valor**	**Significância**	**Entre grupos**
X_{11} Linha do produto	6,388	25,642	0,000	Menos de 1 ano e de 1 a 5 anos

Nota: Em cada passo, a variável que maximiza a distância Mahalanobis entre os dois grupos mais próximos é adicionada.

Variáveis na análise após o passo 6

Variável	Tolerância	*F* para remover	D^2	Entre grupos
X_6 Qualidade do produto	0,754	50,494	0,007	De 1 a 5 anos e mais de 5 anos
X_{18} Velocidade de entrega	0,754	60,646	0,121	Menos de 1 ano e de 1 a 5 anos

Variáveis fora da análise após o passo 6

Variável	Tolerância	Tolerância mínima	*F* para entrar	D^2 mínimo	Entre grupos
X_7 Atividades de comércio eletrônico	0,954	0,728	0,177	6,474	Menos de 1 ano e de 1 a 5 anos
X_8 Suporte técnico	0,999	0,753	0,269	6,495	Menos de 1 ano e de 1 a 5 anos
X_9 Solução de reclamação	0,453	0,349	0,376	6,490	Menos de 1 ano e de 1 a 5 anos
X_{10} Anúncio	0,954	0,742	0,128	6,402	Menos de 1 ano e de 1 a 5 anos
X_{11} Linha do produto	0,701	0,529	2,552	6,916	Menos de 1 ano e de 1 a 5 anos
X_{12} Imagem da equipe de venda	0,957	0,730	0,641	6,697	Menos de 1 ano e de 1 a 5 anos
X_{13} Preços competitivos	0,994	0,749	1,440	6,408	Menos de 1 ano e de 1 a 5 anos
X_{14} Garantia e reclamações	0,991	0,751	0,657	6,694	Menos de 1 ano e de 1 a 5 anos
X_{15} Novos produtos	0,984	0,744	0,151	6,428	Menos de 1 ano e de 1 a 5 anos
X_{16} Encomenda e cobrança	0,682	0,514	2,397	6,750	Menos de 1 ano e de 1 a 5 anos
X_{17} Flexibilidade de preço	0,652	0,628	3,431	6,830	Menos de 1 ano e de 1 a 5 anos

Teste de significância de diferenças de grupos após o passo 6[a]

X_1 Tipo de cliente		Menos de 1 ano	De 1 a 5 anos
De 1 a 5 anos	*F*	25,642	
	Sig.	0,000	
Mais de 5 anos	*F*	110,261	30,756
	Sig.	0,000	0,000

[a]6 e 52 graus de liberdade.

Avaliação da precisão de classificação

Como esse é um modelo de análise discriminante de três grupos, duas funções discriminantes são calculadas para discriminar entre os três grupos. Valores para cada caso são inseridos no modelo discriminante e composições lineares (escores *Z* discriminantes) são calculadas. As funções discriminantes são baseadas somente nas variáveis incluídas no modelo discriminante.

A Tabela 5-19 fornece os pesos discriminantes de ambas as variáveis (X_6 e X_{18}) e as médias de cada grupo em ambas as funções (parte inferior da tabela). Como podemos ver examinando as médias de grupos, a primeira função distingue o grupo 1 (Menos de 1 ano) dos outros dois grupos (apesar de uma sensível diferença ocorrer entre os grupos 2 e 3 também), enquanto a segunda função separa o grupo 3 (Mais de 5 anos) dos outros dois. Portanto, a primeira função fornece a maior separação entre todos os três grupos, mas é complementada pela segunda função, a qual melhor discrimina (1 e 2 versus 3) onde a primeira função é mais fraca.

TABELA 5-19 Estatísticas resumo para análise discriminante de três grupos

Ajuste geral do modelo: funções discriminantes canônicas

		Percentual de variância						
Função	**Autovalor**	**Função (%)**	**Percentual cumulativo**	**Correlação canônica**	**Lambda de Wilks**	**Qui-quadrado**	***df***	**Significância**
1	3,950	91,5	91,5	0,893	0,148	107,932	4	0,000
2	0,365	8,5	100,0	0,517	0,733	17,569	1	0,000

Coeficientes da função discriminante e da função de classificação

	FUNÇÃO DISCRIMINANTE						
	Função discriminante não-padronizada		***Função discriminante padronizada***		***Funções de classificação***		
Variáveis independentes	**Função 1**	**Função 2**	**Função 1**	**Função 2**	**Menos de 1 ano**	**De 1 a 5 anos**	**Acima de 5 anos**
X_{16} Encomenda e cobrança*	0,308	1,159	0,969	0,622	14,382	15,510	18,753
X_{18} Velocidade de entrega	2,200	0,584	1,021	–0,533	25,487	31,185	34,401
(Constante)	–10,832	–11,313			–91,174	–120,351	–159,022

Matriz estrutural

	Cargas discriminantes não-rotacionadas[a]		**Cargas discriminantes rotacionadas[b]**	
Variáveis independentes	**Função 1**	**Função 2**	**Função 1**	**Função 2**
X_9 Solução de reclamação*	0,572	–0,470	0,739	0,039
X_{16} Encomenda e cobrança	0,499	–0,263	0,546	0,143
X_{11} Linha do produto*	0,483	–0,256	0,529	0,137
X_{15} Novos produtos*	0,125	–0,005	0,096	0,080
X_8 Suporte técnico*	0,030	–0,017	0,033	0,008
X_6 Qualidade do produto*	0,463	0,886	–0,257	0,967
X_{18} Velocidade de entrega	0,540	–0,842	0,967	–0,257
X_{17} Flexibilidade de preço*	0,106	–0,580	0,470	–0,356
X_{10} Anúncio*	0,028	–0,213	0,165	–0,138
X_7 Atividades de comércio eletrônico*	–0,095	–0,193	0,061	–0,207
X_{12} Imagem da equipe de venda*	–0,088	–0,188	0,061	–0,198
X_{14} Garantia e reclamações*	0,030	–0,088	0,081	0,044
X_{13} Preços competitivos*	–0,055	–0,059	–0,001	–0,080

[a]Correlações internas de grupos entre variáveis discriminantes e variáveis de funções discriminantes canônicas padronizadas ordenadas por tamanho absoluto da correlação dentro da função.
[b]Correlações internas de grupos entre variáveis discriminantes e funções discriminantes canônicas padronizadas e rotacionadas.
*Esta variável não é usada na análise.

Médias de grupo (centróides) de funções discriminantes[c]

***X_1* Tipo de cliente**	**Função 1**	**Função****
Menos de 1 ano	–1,911	–1,274
De 1 a 5 anos	0,597	–0,968
Mais de 5 anos	1,371	1,625

[c]Funções discriminantes canônicas não-padronizadas avaliadas nas médias de grupos.

Avaliação da precisão preditiva de pertinência a grupo. O passo final para avaliar o ajuste geral do modelo é determinar o nível de precisão preditiva da(s) função(ões) discriminante(s). Essa determinação é conseguida do mesmo modo que se faz no modelo discriminante de dois grupos, examinando-se as matrizes de classificação e o percentual corretamente classificado (razão de sucesso) em cada amostra.

A classificação de casos individuais pode ser executada pelo método de corte descrito no caso de dois grupos

* N. de RT.: Na realidade, foi incluída a variável X_6 (Qualidade do produto).
** N. de RT.: Neste caso, é Função 2.

ou usando as funções de classificação (ver Tabela 5-19) onde cada caso é computado em cada função de classificação e classificado no grupo de maior escore.

A Tabela 5-20 mostra que as duas funções discriminantes em combinação atingem um grau elevado de precisão de classificação. A proporção de sucesso para a amostra de análise é de 86,7%. No entanto, a razão de sucesso para a amostra de teste cai para 55,0%. Esses resultados demonstram o viés ascendente que é típico quando se aplica somente à amostra de análise, mas não a uma amostra de validação.

Ambas as proporções de sucesso devem ser comparadas com os critérios de chance máxima e de chance proporcional para se avaliar sua verdadeira efetividade. O procedimento de validação cruzada é discutido no passo 6.

- O critério de chance máxima é simplesmente a proporção de sucesso obtida se designarmos todas as observações para o grupo com a maior probabilidade de ocorrência. Na presente amostra de 100 observações, 32 estavam no grupo 1, 35 no grupo 2, e 33 no grupo 3. A partir dessa informação, podemos ver que a probabilidade mais alta seria 35% (grupo 2). O valor de referência para a chance máxima (35% × 1,25) é 43,74%.
- O critério de chance proporcional é calculado elevando-se ao quadrado as proporções de cada grupo, com um valor calculado de 33,36% ($0{,}32^2 + 0{,}35^2 + 0{,}33^2 = 0{,}334$) e um valor de referência de 41,7% (33,4% × 1.25 = 41,7%).

As proporções de sucesso para as amostras de análise e de teste (86,7% e 55,0%, respectivamente) excedem ambos os valores de referência de 43,74% e 41,7%. Na amostra de estimação, todos os grupos individuais ultrapassam os dois valores de referência. Na amostra de teste, porém, o grupo 2 tem uma razão de sucesso de somente 40,9%, e aumenta apenas para 53,8% na amostra de análise. Tais resultados mostram que o grupo 2 deve ser o foco no melhoramento da classificação, possivelmente com a adição de variáveis independentes ou uma revisão da classificação de firmas neste grupo para identificar as características do mesmo que não estão representadas na função discriminante.

A medida final de precisão de classificação é o Q de Press, calculado para as amostras de análise e de validação. Ele testa a significância estatística de que a precisão de classificação é melhor do que o acaso (chance).

$$Q \text{ de Press}_{\text{amostra de estimação}} = \frac{[60 - (52 \times 3)]^2}{60(3-1)} = 76{,}8$$

E o cálculo para a amostra de teste é

$$Q \text{ de Press}_{\text{amostra de validação}} = \frac{[40 - (22 \times 3)]^2}{40(3-1)} = 8{,}45$$

(*Continua*)

TABELA 5-20 Resultados de classificação para a análise discriminante de três grupos

Resultados de classificação[a, b, c]

		Pertinência prevista em grupo			
	Grupo real	**Menos do que 1 ano**	**De 1 a 5 anos**	**Mais de 5 anos**	**Total**
Amostra de estimação	Menos de 1 ano	21	1	0	22
		95,5	4,5	0,0	
	De 1 a 5 anos	2	7	4	13
		15,4	53,8	30,8	
	Mais de 5 anos	0	1	24	25
		0,0	4,0	96,0	
Validação cruzada	Menos de 1 ano	21	1	0	22
		95,5	4,5	0,0	
	De 1 a 5 anos	2	7	4	13
		15,4	53,8	30,8	
	Mais de 5 anos	0	1	24	25
		0,0	4,0	96,0	
Amostra de validação	Menos de 1 ano	5	3	2	10
		50,0	30,0	20,0	
	De 1 a 5 anos	1	9	12	22
		4,5	40,9	54,5	
	Mais de 5 anos	0	0	8	8
		0,0	0,0	100,0	

[a]86,7% dos casos agrupados originais selecionados corretamente classificados.
[b]55,0% dos casos agrupados originais não-selecionados corretamente classificados.
[c]86,7% dos casos agrupados selecionados e validados por cruzamento corretamente classificados.

(*Continuação*)

Como o valor crítico em um nível de significância de 0,01 é 6,63, a análise discriminante pode ser descrita como prevendo pertinência a grupo melhor do que o acaso.

Quando completado, podemos concluir que o modelo discriminante é válido e tem níveis adequados de significância estatística e prática para todos os grupos. Os valores consideravelmente menores para a amostra de validação em todos os padrões de comparação, contudo, justificam a preocupação levantada anteriormente sobre as razões de sucesso específicas de grupos e geral.

Diagnósticos por casos

Além das tabelas de classificação mostrando resultados agregados, informação específica de casos também está disponível detalhando a classificação de cada observação. Essa informação pode detalhar as especificidades do processo de classificação ou representar a classificação através de um mapa territorial.

Informação de classificação específica de caso. Uma série de medidas específicas de casos está disponível para identificação dos casos mal classificados, bem como o diagnóstico da extensão de cada classificação ruim. Usando essa informação, padrões entre os mal classificados podem ser identificados.

A Tabela 5-21 contém dados adicionais de classificação para cada caso individual que foi mal classificado (informação similar também está disponível para todos os outros casos, mas foi omitida por problemas de espaço). Os tipos básicos de informação de classificação incluem o que se segue:

- *Pertinência a grupo.* Tanto os grupos reais quanto os previstos são exibidos para identificar cada tipo de má classificação (p.ex., pertinência real ao grupo 1, mas prevista no grupo 2). Neste caso, vemos os 8 casos mal classificados na amostra de análise (verifique acrescentando os valores fora da diagonal na Tabela 5-20) e os 18 casos mal classificados na amostra de validação.
- *Distância de Mahalanobis ao centróide de grupo previsto.* Denota a proximidade desses casos mal classificados em relação ao grupo previsto. Algumas observações, como o caso 10, obviamente são semelhantes às observações do grupo previsto e não do grupo real. Outras observações, como o caso 57 (distância de Mahalanobis de 6,041), são possivelmente observações atípicas no grupo previsto e no grupo real. O mapa territorial discutido na próxima seção retrata graficamente a posição de cada observação e auxilia na interpretação das medidas de distância.
- *Escores discriminantes.* O escore *Z* discriminante para cada caso em cada função discriminante fornece uma maneira de comparação direta entre casos e um posicionamento relativo versus as médias de grupos.
- *Probabilidade de classificação.* Derivada do emprego das funções discriminantes de classificação, a probabilidade de pertinência para cada grupo é dada. Os valores de probabilidade viabilizam ao pesquisador avaliar a extensão da má classificação. Por exemplo, dois casos, 85 e 89, são do mesmo tipo de má classificação (grupo real 2 e grupo previsto 3), mas muito diferentes em suas classificações quando as probabilidades são focadas. O caso 85 representa uma classificação ruim marginal, pois a probabilidade de previsão no grupo real 2 era de 0,462, enquanto no grupo 3 incorretamente previsto ela era um pouco maior (0,529). Esta má classificação contrasta com o caso 89, onde a probabilidade do grupo real era de 0,032, e a probabilidade prevista para o grupo 3 (o mal classificado) era 0,966. Em ambas as situações de má classificação, a extensão ou magnitude varia muito.

O pesquisador deve avaliar a extensão de má classificação para cada caso. Casos que são classificações obviamente ruins devem ser escolhidos para análise adicional (perfil, exame de variáveis adicionais etc.), discutida na análise de dois grupos.

Mapa territorial. A análise de más classificações pode ser suplementada pelo exame gráfico das observações individuais, representando-as com base em seus escores *Z* discriminantes.

A Figura 5-9 mostra cada observação baseada em seus dois escores *Z* discriminantes rotacionados com uma cobertura do mapa territorial que representa as fronteiras dos escores de corte para cada função. Ao ver a dispersão de cada grupo em torno do centróide, podemos observar várias coisas:

- O grupo 3 (Mais de 5 anos) é mais concentrado, com pouca sobreposição com os outros dois grupos, como se mostra na matriz de classificação onde apenas uma observação foi mal classificada (ver Tabela 5-20).
- O grupo 1 (Menos de 1 ano) é o menos compacto, mas o domínio de casos não se sobrepõe em grande grau com os outros grupos, tornando previsões muito melhores do que poderia ser esperado para um grupo tão variado. Os únicos casos mal classificados que são substancialmente distintos são o caso 10, que é próximo ao centróide do grupo 2, e o caso 13, que é próximo ao centróide do grupo 3. Ambos os casos merecem melhor investigação quanto às suas similaridades com outros grupos.
- Estes dois grupos fazem contraste com o grupo 2 (De 1 a 5 anos), que pode ser visto como tendo substancial sobreposição com o grupo 3 e, em menor extensão, com o grupo 1 (Menos de 1 ano). Essa sobreposição resulta nos mais baixos níveis de precisão de classificação nas amostras de análise e de teste.
- A sobreposição que ocorre entre os grupos 2 e 3 no centro e à direita no gráfico sugere a possível existência de um quarto grupo. Uma análise poderia ser levada

(*Continua*)

TABELA 5-21 Previsões mal classificadas para casos individuais na análise discriminante de três grupos

	PERTINÊNCIA A GRUPO			ESCORES DISCRIMINANTES		PROBABILIDADE DE CLASSIFICAÇÃO		
Identificação do caso	*(X_1) Real*	*Previsto*	*Distância de Mahalanobis ao centróide*[a]	*Função 1*	*Função 2*	*Grupo 1*	*Grupo 2*	*Grupo 3*
Amostra de análise/estimação								
10	1	2	0,175	0,81755	–1,32387	0,04173	0,93645	0,02182
8	2	1	1,747	–0,78395	–1,96454	0,75064	0,24904	0,00032
100	2	1	2,820	–0,70077	–0,11060	0,54280	0,39170	0,06550
1	2	3	2,947	–0,07613	0,70175	0,06527	0,28958	0,64515
5	2	3	3,217	–0,36224	1,16458	0,05471	0,13646	0,80884
37	2	3	3,217	–0,36224	1,16458	0,05471	0,13646	0,80884
88	2	3	2,390	0,99763	0,12476	0,00841	0,46212	0,52947
58	3	2	0,727	0,30687	–0,16637	0,07879	0,70022	0,22099
Amostra de teste/validação								
25	1	2	1,723	–0,18552	–2,02118	0,40554	0,59341	0,00104
77	1	2	0,813	0,08688	–0,22477	0,13933	0,70042	0,16025
97	1	2	1,180	–0,41466	–0,57343	0,42296	0,54291	0,03412
13	1	3	0,576	1,77156	2,26982	0,00000	0,00184	0,99816
96	1	3	3,428	–0,26535	0,75928	0,09917	0,27855	0,62228
83	2	1	2,940	–1,58531	0,40887	0,89141	0,08200	0,02659
23	2	3	0,972	0,61462	0,99288	0,00399	0,10959	0,88641
34	2	3	1,717	0,86996	0,41413	0,00712	0,31048	0,68240
39	2	3	0,694	1,59148	0,82119	0,00028	0,08306	0,91667
41	2	3	2,220	0,30230	0,58670	0,02733	0,30246	0,67021
42	2	3	0,210	1,08081	1,97869	0,00006	0,00665	0,99330
55	2	3	1,717	0,86996	0,41413	0,00712	0,31048	0,68240
57	2	3	6,041	3,54521	0,47780	0,00000	0,04641	0,95359
62	2	3	4,088	–0,32690	0,52743	0,17066	0,38259	0,44675
75	2	3	2,947	–0,07613	0,70175	0,06527	0,28958	0,64515
78	2	3	0,210	1,08081	1,97869	0,00006	0,00665	0,99330
85	2	3	2,390	0,99763	0,12476	0,00841	0,46212	0,52947
89	2	3	0,689	0,54850	1,51411	0,00119	0,03255	0,96625

[a]Distância de Mahalanobis ao centróide do grupo previsto

(*Continuação*)

a cabo para determinar o real intervalo de tempo de clientes, talvez com clientes com mais de 1 ano divididos em três grupos ao invés de dois.

A representação gráfica é útil não apenas para identificar esses casos mal classificados que podem formar um novo grupo, mas também para identificar observações atípicas. A discussão anterior indica possíveis opções para identificar observações atípicas (caso 57), bem como a possibilidade de redefinição de grupos entre os grupos 2 e 3.

Estágio 5: Interpretação dos resultados da análise discriminante de três grupos

O próximo estágio da análise discriminante envolve uma série de passos na interpretação das funções discriminantes.

- Calcular as cargas para cada função e rever a rotação das funções para fins de simplificação da interpretação.
- Examinar as contribuições das variáveis preditoras: (a) a cada função separadamente (ou seja, cargas discriminantes), (b) cumulativamente sobre múltiplas funções discriminantes com o índice de potência, e (c) graficamente em uma solução bidimensional para entender a posição relativa de cada grupo e a interpretação das variáveis relevantes na determinação dessa posição.

Cargas discriminantes e suas rotações

Uma vez que as funções discriminantes são calculadas, elas são correlacionadas com todas as variáveis independentes, mesmo aquelas não usadas na função discriminante, para desenvolver uma matriz estrutural (de cargas). Tal procedimento nos permite ver onde a discriminação ocorreria se todas as variáveis independentes fossem incluídas no modelo (ou seja, se nenhuma fosse excluída por multicolinearidade ou falta de significância estatística).

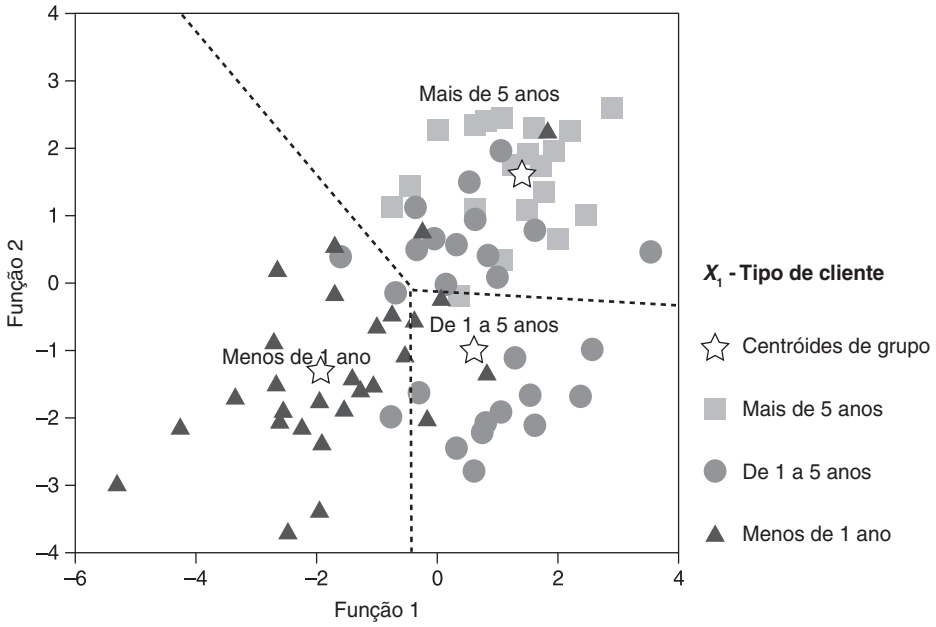

FIGURA 5-9 Mapa territorial para a análise discriminante de três grupos.

Cargas discriminantes. As cargas não-rotacionadas representam a associação de cada variável independente com cada função, mesmo que não esteja incluída na função discriminante. Cargas discriminantes, semelhantes às cargas fatoriais descritas no Capítulo 3, são as correlações entre cada variável independente e o escore discriminante.

A Tabela 5-19 contém a matriz estrutural de cargas não-rotacionadas para ambas as funções discriminantes. Selecionando variáveis com cargas de 0,40 ou acima como descritivas das funções, percebemos que a função 1 tem cinco variáveis excedendo 0,40 (X_9, X_{18}, X_{16}, X_{11} e X_6), enquanto quatro variáveis são descritivas da função 2 (X_6, X_{18}, X_{17} e X_9). Ainda que pudéssemos usar essas variáveis para descrever cada função, enfrentaríamos o problema de que três variáveis (X_9, X_6 e X_{18}) têm cargas duplas (variáveis selecionadas como descritivas de ambas as funções). Se fôssemos proceder com as cargas não-rotacionadas, cada função compartilharia mais variáveis com a outra do que teria feito se fosse única.

A falta de distinção das cargas com cada variável descritiva de uma só função pode ser abordada com rotação da matriz estrutural, exatamente como foi feito com cargas fatoriais. Para uma descrição mais detalhada do processo de rotação, ver Capítulo 3.

Rotação Depois que as cargas da função discriminante são calculadas, elas podem ser rotacionadas para redistribuir a variância (esse conceito é melhor explicado no Capítulo 3). Basicamente, a rotação preserva a estrutura original e a confiabilidade dos modelos discriminantes e facilita consideravelmente a sua interpretação.

Na presente aplicação, escolhemos o procedimento mais amplamente usado de rotação VARIMAX. A rotação afeta os coeficientes da função e as cargas discriminantes, bem como o cálculo dos escores *Z* discriminantes e dos centróides de grupo (ver Tabela 5-19). Examinar os coeficientes ou as cargas rotacionados versus não-rotacionados revela um conjunto de resultados um pouco mais simples (ou seja, as cargas tendem a se separar em valores altos versus baixos, em vez de se limitarem a um domínio intermediário). As cargas rotacionadas permitem interpretações muito mais distintas de cada função:

- A função 1 é agora descrita por três variáveis (X_{18}, X_9 e X_{16}) que formam o fator *Serviço ao Cliente de Pós-Venda* durante a análise fatorial (ver Capítulo 3 para mais detalhes), mais X_{11} e X_{17}. Assim, o serviço a cliente, mais linha de produto e flexibilidade de preço são descritores da função 1.
- A função 2 mostra apenas uma variável, X_6 (Qualidade do produto), que tem uma carga acima de 0,40 para a segunda função. Apesar de X_{17} ter um valor abaixo da referência (–0,356), esta variável tem uma carga maior na primeira função, o que a torna um descritor daquela função. Logo, a segunda função pode ser descrita pela variável de Qualidade do produto.

Com duas ou mais funções estimadas, a rotação pode ser uma poderosa ferramenta que sempre deve ser considerada para aumentar a interpretabilidade dos resultados. Em nosso exemplo, cada uma das variáveis que entrou no

processo *stepwise* era descritiva de uma das funções discriminantes. O que devemos fazer agora é avaliar o impacto de cada variável em termos da análise discriminante geral (i.e., em ambas as funções).

Avaliação da contribuição de variáveis preditoras

Tendo descrito as funções discriminantes em termos das variáveis independentes – tanto aquelas que foram usadas nas funções discriminantes quanto as que não foram incluídas – voltamos nossa atenção para conseguir uma melhor compreensão do impacto das próprias funções, e então das variáveis individuais.

Impacto das funções individuais. A primeira tarefa é examinar as funções discriminantes em termos de como elas diferenciam entre os grupos.

Começamos examinando os centróides de grupos quanto às duas funções como mostrado na Tabela 5-19. Uma abordagem mais fácil é através do mapa territorial (Figura 5-9):

- Examinando os centróides de grupos e a distribuição de casos em cada grupo, percebemos que a função 1 prioritariamente diferencia entre o grupo 1 e os grupos 2 e 3, enquanto a função 2 distingue entre o grupo 3 e os grupos 1 e 2.
- A sobreposição e a má classificação dos casos dos grupos 2 e 3 pode ser tratada via o exame da força das funções discriminantes e dos grupos diferenciados por conta de cada uma. Retomando a Tabela 5-19, a função 1 era, de longe, o discriminador mais potente, e ela prioritariamente separava o grupo 1 dos demais. A função 2, que separava o grupo 3 dos outros, era muito mais fraca em termos de poder discriminante. Não é surpresa que a maior sobreposição e má classificação ocorreriam entre os grupos 2 e 3, que são distinguidos principalmente pela função 2.

Essa abordagem gráfica ilustra as diferenças nos grupos devido às funções discriminantes, mas não fornece uma base para explicar essas diferenças em termos das variáveis independentes.

Para avaliar as contribuições das variáveis individuais, o pesquisador conta com várias medidas – cargas discriminantes, razões *F* univariadas e o índice de potência. As técnicas envolvidas no uso de cargas discriminantes e de razões *F* univariadas foram discutidas no exemplo de dois grupos. Examinaremos mais detalhadamente o índice de potência, um método de avaliação da contribuição de uma variável em múltiplas funções discriminantes.

Índice de potência. O índice de potência é uma técnica adicional de interpretação muito útil em situações com mais de uma função discriminante. Ele retrata a contribuição de cada variável individual em todas as funções discriminantes em termos de uma única medida comparável.

O índice de potência reflete tanto as cargas de cada variável quanto o poder discriminatório relativo de cada função. As cargas rotacionadas representam a correlação entre a variável independente e o escore *Z* discriminante. Assim, a carga ao quadrado é a variância na variável independente associada com a função discriminante. Ponderando a variância explicada de cada função via poder discriminatório relativo da função e somando nas funções, o índice de potência representa o efeito discriminante total de cada variável ao longo de todas as funções discriminantes.

A Tabela 5-22 fornece os detalhes do cálculo do índice de potência para cada variável independente. A comparação das variáveis quanto a seus índices de potência revela o seguinte:

- X_{18} (Velocidade de entrega) é a variável independente responsável pela maior discriminação entre os três tipos de grupos de clientes.
- Ela é seguida em termos de impacto por quatro variáveis não incluídas na função discriminante (X_9, X_{16}, X_{11} e X_{17}).
- A segunda variável na função discriminante (X_6) tem apenas o sexto maior valor de potência.

Por que X_6 tem somente o sexto maior valor de potência mesmo sendo uma das duas variáveis incluídas na função discriminante?

- Primeiro, lembre-se que multicolinearidade afeta soluções *stepwise* devido à redundância entre variáveis altamente multicolineares. X_9 e X_{16} eram as duas variáveis altamente associadas com X_{18} (formando o fator Serviço a Clientes), e assim seu impacto em um sentido univariado, refletido no índice de potência, não era necessário na função discriminante devido à presença de X_{18}.
- As outras duas variáveis, X_{11} e X_{17}, entraram através do procedimento *stepwise*, mas foram removidas uma vez que X_6 foi adicionada, novamente devido à multicolinearidade. Assim, seu maior poder discriminante está refletido em seus valores de potência ainda que elas não fossem necessárias na função discriminante, uma vez que X_6 foi acrescentada com X_{18} na função discriminante.
- Finalmente, X_6, a segunda variável na função discriminante, tem um baixo valor de potência por ser associada com a segunda função discriminante, que tem relativamente pouco impacto discriminante quando comparada com a primeira função. Logo, a despeito de X_6 ser um elemento necessário na discriminação entre os três grupos, seu impacto geral é menor do que aquelas variáveis associadas com a primeira função.

Lembre-se que os valores de potência podem ser calculados para todas as variáveis independentes, mesmo que não estejam nas funções discriminantes, pois eles são baseados em cargas discriminantes. A meta do índice de potência é fornecer interpretação naqueles casos onde

TABELA 5-22 Cálculo dos índices de potência para a análise discriminante de três grupos

	Função discriminante 1				*Função discriminante 2*				
Variáveis independentes	**Carga**	**Carga ao quadrado**	**Autovalor relativo**	**Valor de potência**	**Carga**	**Carga ao quadrado**	**Autovalor relativo**	**Valor de potência**	**Índice de potência**
X_6 Qualidade do produto	–0,257	0,066	0,915	0,060	0,967	0,935	0,085	0,079	0,139
X_7 Atividades de comércio eletrônico	0,061	0,004	0,915	0,056	–0,207	0,043	0,085	0,004	0,060
X_8 Suporte técnico	0,033	0,001	0,915	0,001	0,008	0,000	0,085	0,000	0,001
X_9 Solução de reclamação	0,739	0,546	0,915	0,500	0,039	0,002	0,085	0,000	0,500
X_{10} Anúncio	0,165	0,027	0,915	0,025	–0,138	0,019	0,085	0,002	0,027
X_{11} Linha do produto	0,529	0,280	0,915	0,256	0,137	0,019	0,085	0,002	0,258
X_{12} Imagem da equipe de venda	0,061	0,004	0,915	0,004	–0,198	0,039	0,085	0,003	0,007
X_{13} Preços competitivos	–0,001	0,000	0,915	0,000	–0,080	0,006	0,085	0,001	0,001
X_{14} Garantia e reclamações	0,081	0,007	0,915	0,006	0,044	0,002	0,085	0,000	0,006
X_{15} Novos produtos	0,096	0,009	0,915	0,008	0,080	0,006	0,085	0,001	0,009
X_{16} Encomenda e cobrança	0,546	0,298	0,915	0,273	0,143	0,020	0,085	0,002	0,275
X_{17} Flexibilidade de preço	0,470	0,221	0,915	0,202	–0,356	0,127	0,085	0,011	0,213
X_{18} Velocidade de entrega	0,967	0,935	0,915	0,855	–0,257	0,066	0,085	0,006	0,861

Nota: O autovalor relativo de cada função discriminante é calculado como o autovalor de cada função (mostrado na Tabela 5-19 como 3,950 e 0,365 para as funções discriminantes I e II, respectivamente) dividido pelo total dos autovalores (3,950 + 0,365 = 4,315).

multicolinearidade ou outros fatores possam ter evitado a inclusão de uma variável na função discriminante.

Uma visão geral das medidas empíricas de impacto. Como visto nas discussões anteriores, o poder discriminatório de variáveis em análise discriminante é refletido em muitas medidas diferentes, cada uma desempenhando um papel único na interpretação dos resultados discriminantes. Combinando todas essas medidas em nossa avaliação das variáveis, podemos conquistar uma perspectiva bastante eclética sobre como cada variável se ajusta nos resultados discriminantes.

A Tabela 5-23 apresenta as três medidas interpretativas preferidas (cargas rotacionadas, razão *F* univariada e índice de potência) para cada variável independente. Os resultados apóiam a análise *stepwise*, apesar de ilustrarem em diversos casos o impacto de multicolinearidade sobre os procedimentos e os resultados.

- Duas variáveis (X_9 e X_{18}) têm os maiores impactos individuais como evidenciado por seus valores *F* univariados. No entanto, como ambas são altamente associadas (como evidenciado por suas inclusões no fator Serviço ao cliente do Capítulo 3), apenas uma será incluída em uma solução *stepwise*. Ainda que X_9 tenha um valor *F* univariado marginalmente maior, a habilidade de X_{18} fornecer uma melhor discriminação entre todos os grupos (como evidenciado por seu maior valor mínimo D^2 de Mahalanobis descrito anteriormente) fez dela a melhor candidata para inclusão. Portanto, X_9, em uma base individual, tem um poder discriminante comparável, mas X_{18} será vista funcionando melhor com outras variáveis.
- Três variáveis adicionais (X_6, X_{11} e X_{16}) são as próximas com maior impacto, mas apenas uma, X_6, é mantida na função discriminante. Note que X_{16} é altamente correlacionada com X_{18} (ambas parte do fator Serviço ao cliente) e não incluída na função discriminante, enquanto X_{11} entrou na mesma, mas foi uma daquelas variáveis removidas depois que X_6 foi adicionada.
- Finalmente, duas variáveis (X_{17} e X_{13}) tinham quase os mesmos efeitos univariados, mas somente X_{17} tinha uma associação substancial com uma das funções discriminantes (uma carga de 0,470 sobre a primeira função). O resultado é que mesmo que X_{17} possa ser considerada descritiva da primeira função e tendo um impacto na discriminação baseado nessas funções, X_{13} não tem qualquer impacto, seja em associação com essas duas funções, seja em adição uma vez que estas funções sejam explicadas.
- Todas as variáveis remanescentes tinham pequenos valores *F* univariados e pequenos valores de potência, o que indica pouco ou nenhum impacto tanto no sentido univariado quanto multivariado.

De particular interesse é a interpretação das duas dimensões de discriminação. Essa interpretação pode ser feita somente através do exame das cargas, mas é complementada por uma representação gráfica das cargas discriminantes, como descrito na próxima seção.

Representação gráfica de cargas discriminantes. Para representar as diferenças em termos das variáveis preditoras, as cargas e os centróides de grupos podem ser representados graficamente em espaço discriminante reduzido. Como observado anteriormente, a representação mais válida é o uso de vetores de atribuição e centróides de grupos expandidos.

A Tabela 5-24 mostra os cálculos para a expansão das cargas discriminantes (usadas para vetores de atribuição) e de centróides de grupos. O processo de represen-

(*Continua*)

TABELA 5-23 Resumo de medidas interpretativas para análise discriminante de três grupos

		Cargas rotacionadas de função discriminante			
		Função 1	**Função 2**	**Razão *F* univariada**	**Índice de potência**
X_6	Qualidade do produto	–0,257	0,967	32,311	0,139
X_7	Atividades de comércio eletrônico	0,061	–0,207	1,221	0,060
X_8	Suporte técnico	0,033	0,008	0,782	0,001
X_9	Solução de reclamação	0,739	0,039	40,292	0,500
X_{10}	Anúncio	0,165	–0,138	1,147	0,027
X_{11}	Linha do produto	0,529	0,137	32,583	0,258
X_{12}	Imagem da equipe de venda	0,061	–0,198	1,708	0,007
X_{13}	Preços competitivos	–0,001	–0,080	9,432	0,001
X_{14}	Garantia e reclamações	0,081	0,044	2,619	0,006
X_{15}	Novos produtos	0,096	0,080	0,216	0,009
X_{16}	Encomenda e cobrança	0,546	0,143	25,048	0,275
X_{17}	Flexibilidade de preço	0,470	–0,356	12,551	0,213
X_{18}	Velocidade de entrega	0,967	–0,257	40,176	0,861

(*Continuação*)

tação gráfica sempre envolve todas as variáveis incluídas no modelo pelo procedimento *stepwise* (em nosso exemplo, X_6 e X_{18}). No entanto, também faremos o gráfico das variáveis não incluídas na função discriminante se suas respectivas razões *F* univariadas forem significantes, o que adiciona X_9, X_{11} e X_{16} ao espaço discriminante reduzido. Esse procedimento mostra a importância de variáveis colineares que não foram incluídas no modelo *stepwise* final, semelhante ao índice de potência.

Os gráficos dos vetores de atribuição expandidos para as cargas discriminantes rotacionadas são exibidos na Figura 5-10. Os vetores do gráfico nos quais esse procedimento foi usado apontam para os grupos que têm a mais alta média na respectiva variável independente e para a direção oposta dos grupos que têm os mais baixos escores médios. Assim, a interpretação do gráfico na Figura 5-10 indica o seguinte:

- Como observado no mapa territorial e na análise dos centróides de grupos, a primeira função discriminante distingue entre grupo 1 e grupos 2 e 3, enquanto a segunda diferencia o grupo 3 dos grupos 1 e 2.
- A correspondência de X_{11}, X_{16}, X_9 e X_{18} com o eixo X reflete a associação delas com a primeira função discriminante, enquanto vemos que somente X_6 é associada com a segunda função discriminante. A figura ilustra graficamente as cargas rotacionadas para cada função e distingue as variáveis descritivas de cada função.

Estágio 6: Validação dos resultados discriminantes

As razões de sucesso para as matrizes de classificação cruzada e de teste podem ser usadas para avaliar a validade interna e externa, respectivamente, da análise discriminante. Se as razões de sucesso excederem os valores de referência nos padrões de comparação, então validade é estabelecida. Como anteriormente descrito, os valores de referência são 41,7% para o critério de chance proporcional e 43,7% para o critério de chance máxima. Os resultados de classificação mostrados na Tabela 5-20 fornecem o seguinte suporte para validade:

Validade interna é avaliada pelo método de classificação cruzada, onde o modelo discriminante é estimado deixando um caso de fora e então prevendo aquele caso com o modelo estimado. Este processo é feito em turnos para cada observação, de modo que uma observação jamais influencia o modelo discriminante que prevê sua classificação em algum grupo.

Como visto na Tabela 5-20, a razão de sucesso geral para o método de classificação cruzada de 86,7 substancialmente excede ambos os padrões, tanto geral quanto para cada grupo. Contudo, ainda que todos os três grupos também tenham razões individuais de sucesso acima dos padrões, a razão de sucesso do grupo 2 (53,8) é consideravelmente menor do que aquela sobre os outros dois grupos.

TABELA 5-24 Cálculo dos vetores de atribuição e dos centróides de grupos expandidos no espaço discriminante reduzido

	Cargas da função discriminante rotacionada			*Coordenadas no espaço reduzido*	
Variáveis independentes	**Função 1**	**Função 2**	**Razão *F* univariada**	**Função 1**	**Função 2**
X_6 Qualidade do produto	–0,257	0,967	32,311	–8,303	31,244
X_7 Atividades de comércio eletrônico[a]	0,061	–0,207	1,221		
X_8 Suporte técnico[a]	0,033	0,008	0,782		
X_9 Solução de reclamação	0,739	0,039	40,292	29,776	1,571
X_{10} Anúncio[a]	0,165	–0,138	1,147		
X_{11} Linha do produto	0,529	0,137	32,583	17,236	4,464
X_{12} Imagem da equipe de venda[a]	0,061	–0,198	1,708		
X_{13} Preços competitivos[a]	–0,001	–0,080	9,432		
X_{14} Garantia e reclamações[a]	0,081	0,044	2,619		
X_{15} Novos produtos[a]	0,096	0,080	0,216		
X_{16} Encomenda e cobrança	0,546	0,143	25,048	13,676	3,581
X_{17} Flexibilidade de preço[a]	0,470	–0,356	12,551		
X_{18} Velocidade de entrega	0,967	–0,257	40,176	38,850	–10,325

[a]Variáveis com razões univariadas não-significantes não são representadas no espaço reduzido.

	Centróides de grupo		*Valor F aproximado*		*Coordenadas no espaço reduzido*	
	Função 1	**Função 2**	**Função 1**	**Função 2**	**Função 1**	**Função 2**
Grupo 1: Menos de 1 ano	–1,911	–1,274	66,011	56,954	–126,147	–72,559
Grupo 2: De 1 a 5 anos	0,597	–0,968	66,011	56,954	39,408	–55,131
Grupo 3: Mais de 5 anos	1,371	1,625	66,011	56,954	90,501	92,550

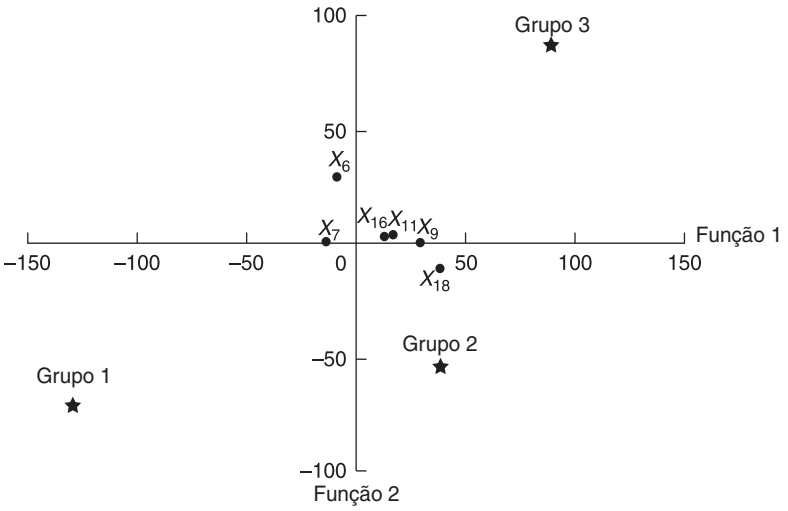

FIGURA 5-10 Gráfico de vetores de atribuição expandidos (variáveis) no espaço discriminante reduzido.

Validade externa é tratada através da amostra de teste, a qual é uma amostra completamente separada que utiliza as funções discriminantes estimadas com a amostra de análise para previsão de grupos.

Em nosso exemplo, a amostra de teste tem uma razão geral de sucesso de 55,0%, o que excede ambos os valores de referência, apesar de isso não ocorrer na magnitude encontrada na abordagem de classificação cruzada. O grupo 2, contudo, não excedeu qualquer valor de referência. Quando as classificações ruins são analisadas, percebemos que mais casos são mal classificados no grupo 3 do que corretamente classificados no grupo 2, o que sugere que esses casos mal classificados sejam examinados diante da possibilidade de uma redefinição dos grupos 2 e 3 para que se crie um novo grupo.

O pesquisador também é encorajado a estender o processo de validação por meio do perfil dos grupos quanto a conjuntos adicionais de variáveis ou aplicando a função discriminante em outra(s) amostra(s) representativa(s) da população geral ou de segmentos da mesma. Além disso, a análise de casos mal classificados ajudará a estabelecer se são necessárias variáveis adicionais ou se a classificação de grupos dependentes precisa de revisão.

Uma visão gerencial

A análise discriminante teve por meta entender as diferenças perceptuais de clientes com base nos intervalos de tempo como clientes da HBAT. Espera-se que o exame de diferenças em percepções HBAT baseadas na constância como clientes identifique percepções que são críticas ao desenvolvimento de uma relação de clientela, o que é tipificado por aqueles clientes de longo prazo. Três grupos de clientela foram formados – menos de 1 ano, de 1 a 5 anos, e mais de 5 anos – e as percepções quanto à HBAT foram medidas sobre 13 variáveis. A análise produziu diversas descobertas importantes, tanto em termos dos tipos de variáveis que distinguiam entre os grupos quanto nos padrões de mudanças ao longo do tempo:

- Primeiro, há duas dimensões de discriminação entre os três grupos de clientes. A primeira dimensão é tipificada por elevadas percepções de serviço aos clientes (Solução de reclamação, Velocidade de entrega e Encomenda e cobrança), juntamente com Linha do produto e Flexibilidade de preço. Em contraste, a segunda dimensão é caracterizada somente em termos de Qualidade do produto.
- O perfil dos três grupos quanto a essas duas dimensões e variáveis associadas com cada dimensão permite à gerência compreender as diferenças perceptuais entre eles.
  - O grupo 1, clientes há menos de 1 ano, geralmente tem as menores percepções da HBAT. Para as três variáveis de serviço à clientela (Solução de reclamação, Encomenda e cobrança, e Velocidade de entrega), esses clientes têm percepções menores do que em qualquer outro grupo. Para Qualidade de produto, Linha de produto e Preço competitivo, este grupo é comparável com o 2 (de 1 a 5 anos), mas ainda tem percepções menores do que clientes há mais de 5 anos. Somente para Flexibilidade de preço este grupo é comparável com os clientes mais antigos e ambos têm valores menores do que os clientes de 1 a 5 anos. No geral, as percepções desses clientes mais recentes seguem o padrão esperado de serem menores do que outros da clientela, mas é esperado

(*Continua*)

(*Continuação*)

que melhorem à medida que permanecerem clientes ao longo do tempo.

- O grupo 2, clientes de 1 a 5 anos, tem semelhanças com os clientes mais novos e os mais antigos. Quanto às três variáveis de serviço à clientela, eles são comparáveis ao grupo 3 (mais de 5 anos). Para Qualidade de produto, Linha de produto e Preço competitivo, suas percepções são mais comparáveis com as dos clientes mais novos (e menores do que as dos clientes mais antigos). Eles mantêm as mais elevadas percepções, dos três grupos, quanto à Flexibilidade de preço.
- O grupo 3, representando os clientes há mais de 5 anos, tem as mais favoráveis percepções da HBAT, como o esperado. Apesar de serem comparáveis aos clientes do grupo 2 quanto às três variáveis de serviço à clientela (com ambos os grupos maiores do que o grupo 1), eles são significantemente maiores que os clientes nos outros dois grupos em termos de Qualidade de produto, Linha de produto e Preço competitivo. Assim, este grupo representa aqueles clientes que têm percepções positivas e têm progredido no estabelecimento de uma relação cliente/HBAT através de um fortalecimento de suas percepções.

- Usando os três grupos como indicadores no desenvolvimento de relações de clientela, podemos identificar dois estágios nos quais as percepções HBAT mudam nesse processo de desenvolvimento:
  - *Estágio 1:* O primeiro conjunto de percepções a mudar é aquele relacionado a serviços a clientes (visto nas diferenças entre os grupos 1 e 2). Este estágio reflete a habilidade da HBAT de afetar positivamente percepções com operações relativas a serviços.
  - *Estágio 2:* Um desenvolvimento de maior prazo é necessário para promover melhoras em elementos mais centrais (Qualidade de produto, Linha de produto e Preço competitivo). Quando ocorrem essas mudanças, o cliente deve se tornar mais comprometido com a relação, como se evidencia por uma longa permanência com a HBAT.
- Deve ser observado que existe evidência de que vários clientes fazem a transição através do estágio 2 mais rapidamente do que os cinco anos, como mostrado pelo considerável número de clientes que têm sido do grupo entre 1 e 5 anos, ainda que mantenham as mesmas percepções da clientela mais antiga. Assim, HBAT pode esperar que certos clientes possam se deslocar através desse processo muito rapidamente, e uma análise mais detalhada sobre tais clientes pode identificar características que facilitam o desenvolvimento de relações com a clientela.

Assim, o gerenciamento leva em conta um *input* para planejamento estratégico e tático não apenas dos resultados diretos da análise discriminante, mas também dos erros de classificação.

REGRESSÃO LOGÍSTICA: REGRESSÃO COM UMA VARIÁVEL DEPENDENTE BINÁRIA

Como discutimos, a análise discriminante é apropriada quando a variável dependente é não-métrica. No entanto, quando a variável dependente tem apenas dois grupos, a regressão logística pode ser preferida por duas razões:

- A análise discriminante depende estritamente de se atenderem as suposições de normalidade multivariada e de igualdade entre as matrizes de variância-covariância nos grupos – suposições que não são atendidas em muitas situações. A regressão logística não depende dessas suposições rígidas e é muito mais robusta quando tais pressupostos não são satisfeitos, o que torna sua aplicação apropriada em muitas situações.
- Mesmo quando os pressupostos são satisfeitos, muitos pesquisadores preferem a regressão logística por ser similar à regressão múltipla. Ela tem testes estatísticos diretos, tratamentos similares para incorporar variáveis métricas e não-métricas e efeitos não-lineares, e uma vasta gama de diagnósticos.

Por essas e outras razões mais técnicas, a regressão logística é equivalente à análise discriminante de dois grupos e pode ser mais adequada em muitas situações.

Nossa discussão de regressão logística não cobre cada um dos seis passos do processo de decisão, mas destaca as diferenças e semelhanças entre a regressão logística e a análise discriminante ou a regressão múltipla. Para uma revisão completa de regressão múltipla, ver o Capítulo 4.

Representação da variável dependente binária

Em análise discriminante, o caráter não-métrico de uma variável dependente dicotômica é acomodado fazendo-se previsões de pertinência a grupo baseadas em escores Z discriminantes. Isso requer o cálculo de escores de corte e a designação de observações a grupos.

A regressão logística aborda essa tarefa de uma maneira mais semelhante à encontrada em regressão múltipla. Regressão logística representa os dois grupos de interesse como uma variável binária com valores de 0 e 1. Não importa qual grupo é designado com o valor de 1 versus 0, mas tal designação deve ser observada para a interpretação dos coeficientes.

- Se os grupos representam características (p.ex., sexo), então um grupo pode ser designado com o valor 1 (p.ex., feminino) e o outro grupo com o valor 0 (p.ex., masculino). Em tal situação, os coeficientes refletiriam o impacto das variáveis independentes sobre a probabilidade da pessoa ser do sexo feminino (ou seja, o grupo codificado como 1).
- Se os grupos representam resultados ou eventos (p.ex., sucesso ou fracasso, compra ou não-compra), a designação dos códigos de grupos causa impacto na interpretação também. Considere que o grupo com sucesso é codificado como 1, e aquele com fracasso, como 0. Então, os coeficientes repre-

sentam os impactos sobre a probabilidade de sucesso. De maneira igualmente fácil, os códigos poderiam ser invertidos (1 agora denota fracasso) e os coeficientes representariam as forças que aumentam a probabilidade de fracasso.

A regressão logística difere da regressão múltipla, contudo, no sentido de que ela foi especificamente elaborada para prever a probabilidade de um evento ocorrer (ou seja, a probabilidade de uma observação estar no grupo codificado como 1). Apesar de os valores de probabilidade serem medidas métricas, há diferenças fundamentais entre regressão múltipla e logística.

Uso da curva logística

Como a variável dependente tem apenas os valores 0 e 1, o valor previsto (probabilidade) deve ser limitado para cair dentro do mesmo domínio. Para definir uma relação limitada por 0 e 1, a regressão logística usa a **curva logística** para representar a relação entre as variáveis independentes e dependente (ver Figura 5-11). Em níveis muito baixos da variável independente, a probabilidade se aproxima de 0, mas nunca alcança tal valor. Analogamente, quando a variável independente aumenta, os valores previstos crescem para acima da curva, mas em seguida a inclinação começa a diminuir de modo que em qualquer nível da variável independente a probabilidade se aproximará de 1,0, mas jamais excederá tal valor. Como vimos em nossas discussões sobre regressão, no Capítulo 4, os modelos lineares de regressão não podem acomodar tal relação, já que ela é inerentemente não-linear. A relação linear de regressão, mesmo com termos adicionais de transformações para efeitos não-lineares, não pode garantir que os valores previstos permaneçam no intervalo de 0 a 1.

Natureza única da variável dependente

A natureza binária da variável dependente (0 ou 1) tem propriedades que violam as suposições da regressão múltipla. Primeiro, o termo de erro de uma variável discreta segue a distribuição binomial ao invés da normal, invalidando assim todos os testes estatísticos que se sustentam nas suposições de normalidade. Segundo, a variância de uma variável dicotômica não é constante, criando casos de heteroscedasticidade também. Além disso, nenhuma violação pode ser remediada por meio de transformações das variáveis dependente ou independentes.

A regressão logística foi desenvolvida para lidar especificamente com essas questões. Não obstante, sua relação única entre variáveis dependente e independentes exige uma abordagem um tanto diferente para estimar a variável estatística, avaliar adequação de ajuste e interpretar os coeficientes, quando comparada com regressão múltipla.

Estimação do modelo de regressão logística

A regressão logística tem uma única variável estatística composta de coeficientes estimados para cada variável independente – como na regressão múltipla. Tal variável estatística é estimada de uma maneira diferente. A regressão logística deriva seu nome da **transformação logit** usada com a variável dependente, criando diversas diferenças no processo de estimação (bem como o processo de interpretação discutido na próxima seção).

Transformação da variável dependente

Como mostrado anteriormente, o modelo logit usa a forma específica da curva logística, que é em forma de S para ficar no domínio de 0 a 1. Para estimar um modelo de regressão logística, essa curva de valores previstos é ajustada aos dados reais, exatamente como foi feito com uma relação linear em regressão múltipla. No entanto, como os valores reais dos dados das variáveis dependentes podem ser somente 0 ou 1, o processo é de algum modo diferente.

A Figura 5-12 retrata dois exemplos hipotéticos de ajuste de uma relação logística aos dados da amostra. Os dados reais representam se um evento acontece ou não designando valores 1 ou 0 aos resultados (neste caso 1 é designado quando o evento ocorreu, 0 no caso contrário,

(*Continua*)

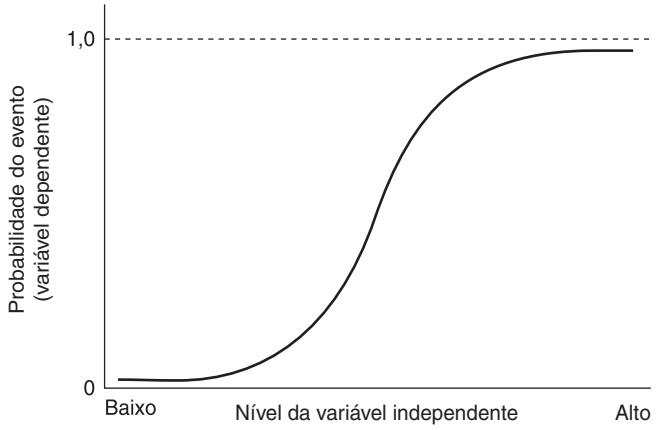

FIGURA 5-11 Forma da relação logística entre variáveis dependente e independentes.

(*Continuação*)

mas tal atribuição poderia facilmente ser invertida). Observações são representadas pelos pontos no topo ou na base do gráfico. Esses resultados (que aconteceram ou não) ocorrem em cada valor da variável independente (o eixo *X*). Na parte (a), a curva logística não pode ajustar bem os dados porque há diversos valores da variável independente que têm ambos os resultados (1 e 0). Neste caso, a variável independente não distingue entre os dois resultados, como se mostra na considerável sobreposição dos dois grupos.

No entanto, na parte (b), uma relação muito melhor definida está baseada na variável independente. Valores menores da variável independente correspondem às observações com 0 para a variável dependente, enquanto valores maiores correspondem bem àquelas observações com um valor 1 sobre a variável dependente. Assim, a curva logística deve ser capaz de ajustar bem os dados.

Mas como prevemos pertinência a grupo a partir da curva logística? Para cada observação, a técnica de regressão logística prevê um valor de probabilidade entre 0 e 1. O gráfico dos valores previstos para todos os valores da variável independente gera a curva exibida na Figura 5-12. Tal probabilidade prevista é baseada nos valores das variáveis independentes e nos coeficientes estimados. Se a probabilidade prevista é maior do que 0,50, então a previsão é de que o resultado seja 1 (o evento ocorreu); caso contrário, o resultado é previsto como sendo 0 (o evento não ocorreu). Retornemos ao nosso exemplo para ver como isso funciona.

Nas partes (a) e (b) da Figura 5-12, um valor de 6,0 para *X* (a variável independente) corresponde a uma probabilidade de 0,50. Na parte (a), podemos ver que diversas observações de ambos os grupos recaem em ambos os lados deste valor, resultando em diversas classificações

(*Continua*)

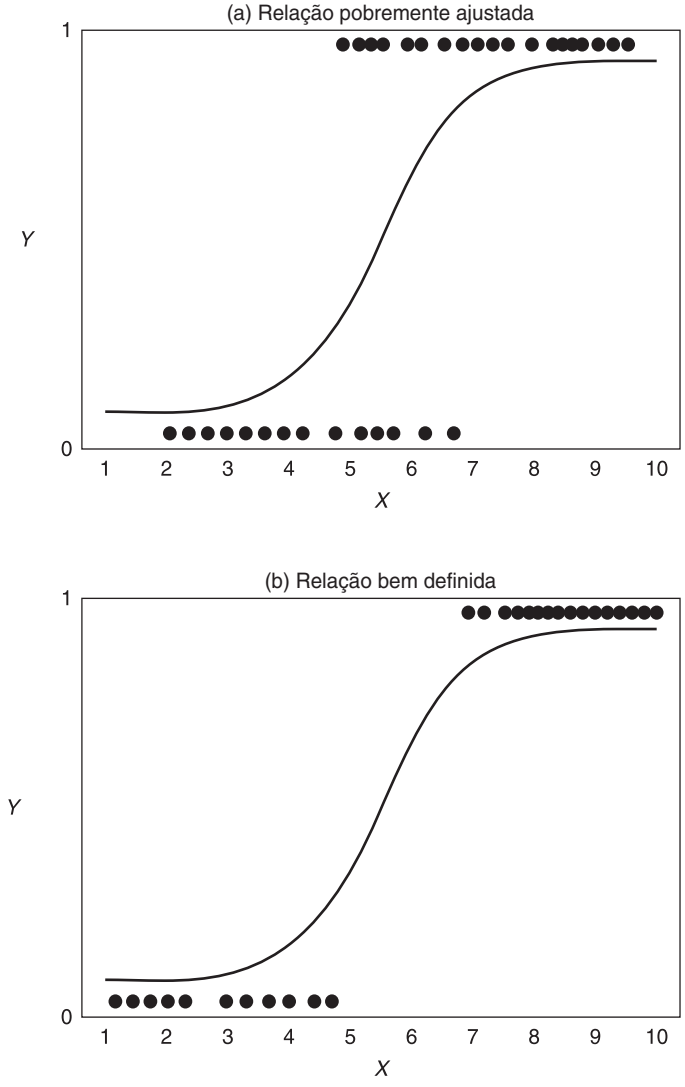

FIGURA 5-12 Exemplos de ajuste da curva logística aos dados da amostra.

(*Continuação*)
ruins. As classificações ruins são mais perceptíveis para o grupo com valores 1,0, ainda que diversas observações no outro grupo (variável dependente = 0,0) também sejam mal classificadas. Na parte (b), fazemos classificação perfeita dos dois grupos quando usamos o valor de probabilidade de 0,50 como valor de corte.

Logo, com uma curva logística estimada, podemos estimar a probabilidade para qualquer observação com base em seus valores para as variáveis independentes e então prever a pertinência a grupo usando 0,50 como valor de corte. Uma vez que temos a pertinência prevista, podemos criar uma matriz de classificação exatamente como foi feito em análise discriminante e avaliar a precisão preditiva.

Estimação dos coeficientes

De onde vem a curva? Em regressão múltipla, estimamos uma relação linear que melhor ajusta os dados. Em regressão logística, seguimos o mesmo processo de previsão da variável dependente por uma variável estatística composta dos **coeficientes logísticos** e as correspondentes variáveis independentes. No entanto, o que difere é que em regressão logística os valores previstos jamais podem estar fora do domínio de 0 a 1. Apesar de uma discussão completa sobre os aspectos conceituais e estatísticos envolvidos no processo de estimação estar além do escopo deste texto, diversas fontes excelentes com tratamentos completos sobre tais aspectos estão disponíveis [3,15,17]. Podemos descrever o processo de estimação em dois passos básicos à medida que introduzimos alguns termos comuns e fornecemos uma breve visão geral do processo.

Transformação de uma probabilidade em razão de desigualdade e valores logit. Como na regressão múltipla, a regressão logística prevê uma variável dependente métrica, neste caso valores de probabilidade restritos ao domínio entre 0 e 1. Mas como podemos garantir que valores estimados não recaiam fora desse domínio? A transformação logística perfaz este processo em dois passos.

Reestabelecimento de uma probabilidade como razão de desigualdades. Em sua forma original, probabilidades não são restritas a valores entre 0 e 1. Portanto, o que aconteceria se reestabelecêssemos a probabilidade de uma maneira que a nova variável sempre ficasse entre 0 e 1? Fazemos isso expressando uma probabilidade como razão de **desigualdades** – a razão entre as probabilidades dos dois resultados ou eventos, $\text{Prob}_i/(1 - \text{Prob}_i)$. Desta forma, qualquer valor de probabilidade é agora dado em uma variável métrica que pode ser diretamente estimada. Qualquer razão de desigualdade pode ser convertida reciprocamente em uma probabilidade que fica entre 0 e 1. Resolvemos nosso problema de restrição dos valores previstos entre 0 e 1 prevendo a razão de desigualdades e então convertendo a mesma em uma probabilidade.

Usemos alguns exemplos da probabilidade de sucesso ou fracasso para ilustrar como a razão de desigualdades é calculada. Se a probabilidade de sucesso é 0,80, então sabemos também que a probabilidade do resultado alternativo (ou seja, o fracasso) é 0,20 (0,20 = 1,0 – 0,80). Esta probabilidade significa que as desigualdades de sucesso são 4,0 (0,80/0,20), ou que o sucesso é quatro vezes mais provável de acontecer do que o fracasso. Reciprocamente, podemos estabelecer as desigualdades de fracasso como 0,25 (0,20/0,80), ou, em outras palavras, o fracasso acontece a um quarto da taxa de sucesso. Assim, qualquer que seja o resultado que busquemos (sucesso ou fracasso), podemos estabelecer a probabilidade como uma chance ou uma razão de desigualdades.

Como você provavelmente já desconfiou, uma probabilidade de 0,50 resulta em razão de desigualdades de 1,0 (ambos os resultados têm iguais chances de ocorrerem). Razão de desigualdades inferior a 1,0 representa probabilidades menores do que 0,50, e razão de desigualdades maior do que 1,0 corresponde a uma probabilidade maior do que 0,50. Agora temos uma variável métrica que sempre pode ser convertida de volta a uma probabilidade entre 0 e 1.

Cálculo do valor logit. A variável de razão de desigualdades resolve o problema de fazer estimativas de probabilidade entre 0 e 1, mas temos outro problema: como fazemos com que as razões de desigualdades fiquem abaixo de 0, que é o limite inferior (não há limite superior). A solução é computar aquilo que é chamado de *valor logit* – calculado via logaritmo das razões de desigualdades. Razões menores que 1,0 têm um logit negativo, razões maiores que 1,0 têm valores logit positivos, e a razão de desigualdades igual a 1,0 (correspondente a uma probabilidade de 0,5) tem um valor logit de 0. Além disso, não importa o quão baixo o valor negativo fique, ele ainda pode ser transformado tomando-se o anti-logaritmo em uma razão de desigualdades maior que 0. O que se segue mostra alguns valores típicos de probabilidade e as razões de desigualdades correspondentes, bem como valores logarítmicos.

Probabilidade	Razão de desigualdades	Logaritmo (Logit)
0,00	0,00	NC
0,10	0,111	–2,197
0,30	0,428	–0,847
0,50	1,000	0,000
0,70	2,333	0,847
0,90	9,000	2,197
1,00	NC	NC

NC = Não pode ser calculado

Com o valor logit, agora temos uma variável métrica que pode ter valores positivos e negativos, mas que sempre pode ser transformada de volta em um valor de probabilidade entre 0 e 1. Observe, no entanto, que o logit jamais pode realmente alcançar 0 ou 1. Esse valor agora se torna a variável dependente do modelo de regressão logística.

Estimação do modelo. Uma vez que compreendemos como interpretar os valores das razões de desigualdades ou das medidas logit, podemos proceder com o uso delas como medida dependente em nossa regressão logística. O processo de estimação dos coeficientes logísticos é semelhante àquele usado em regressão, apesar de que neste caso somente dois valores reais são empregados para a variável dependente (0 e 1). Além do mais, em vez de usar os mínimos quadrados ordinários como meio para estimar o modelo, o método de verossimilhança máxima é utilizado.

Estimação dos coeficientes. Os coeficientes estimados para as variáveis independentes são estimados usando-se o valor logit ou a razão de desigualdades como medida dependente. Cada uma dessas formulações de modelo é exibida aqui:

$$Logit_i = \ln\left(\frac{prob_{evento}}{1 - prob_{evento}}\right) = b_0 + b_1X_1 + ... + b_nX_n$$

ou

$$\text{Razão de desigualdades}_i = \left(\frac{prob_{\text{evento}}}{1 - prob_{\text{evento}}}\right) = e^{b_0 + b_1X_1 + \cdots + b_nX_n}$$

Ambas as formulações de modelo são equivalentes, mas aquela que for escolhida afetará a estimação dos coeficientes. Muitos programas de computador fornecem os coeficientes logísticos em ambas as formas, de modo que o pesquisador deve entender como interpretar cada forma. Discutimos aspectos interpretativos em uma seção posterior.

Este processo pode acomodar uma ou mais variáveis independentes, e estas podem ser métricas ou não-métricas (binárias). Como vemos adiante em nossa discussão sobre interpretação dos coeficientes, ambas as formas dos mesmos refletem a direção e a magnitude da relação, mas são interpretadas de maneiras distintas.

Uso da máxima verossimilhança para estimação. Regressão múltipla emprega o método de mínimos quadrados, que minimiza a soma das diferenças quadradas entre os valores reais e previstos da variável dependente. A natureza não-linear da transformação logística requer que outro procedimento, o da máxima verossimilhança, seja usado de maneira iterativa para que se encontrem as estimativas mais prováveis para os coeficientes. No lugar de minimizar os desvios quadrados (mínimos quadrados), a regressão logística maximiza a probabilidade de que um evento ocorra. O valor de probabilidade, ao invés da soma de quadrados, é em seguida usado quando se calcula uma medida de ajuste geral do modelo. Usar esta técnica alternativa de estimação também demanda que avaliemos o ajuste do modelo de diferentes maneiras.

Avaliação da qualidade do ajuste do modelo de estimação

A qualidade de ajuste para um modelo de regressão logística pode ser avaliada de duas maneiras. Uma é a avaliação de ajuste usando valores "pseudo" R^2, semelhantes àqueles encontrados em regressão múltipla. A segunda abordagem é examinar precisão preditiva (como a matriz de classificação em análise discriminante). As duas técnicas examinam ajuste de modelo sob diferentes perspectivas, mas devem conduzir a conclusões semelhantes.

Ajuste de estimação do modelo

A medida básica do quão bem o procedimento de estimação de máxima verossimilhança se ajusta é o **valor de verossimilhança**, semelhante aos valores das somas de quadrados usadas em regressão múltipla. Regressão logística mede o ajuste da estimação do modelo com o valor –2 vezes o logaritmo do valor da verossimilhança, chamado de $-2LL$ ou –2log verossimilhança. O valor mínimo para $-2LL$ é 0, o que corresponde a um ajuste perfeito (verossimilhança = 1 e $-2LL$ é então 0). Assim, quanto menor o valor $-2LL$, melhor o ajuste do modelo. Como será discutido na próxima seção, o valor $-2LL$ pode ser usado para comparar equações quanto à variação no ajuste ou ser utilizado para calcular medidas comparáveis ao R^2 em regressão múltipla.

Entre comparações de modelos. O valor de verossimilhança pode ser comparado entre equações para avaliar a diferença em ajuste preditivo de uma equação para outra, com testes estatísticos para a significância dessas diferenças. O método básico segue três passos:

1. *Estimar um modelo nulo.* O primeiro passo é calcular um modelo nulo, que atua como a referência para fazer comparações de melhoramento no ajuste do modelo. O modelo nulo mais comum é um sem variáveis independentes, que é semelhante a calcular a soma total de quadrados usando somente a média em regressão múltipla. A lógica por trás desta forma de modelo nulo é que ele pode atuar como uma referência em relação à qual qualquer modelo contendo variáveis independentes pode ser comparado.
2. *Estimar o modelo proposto.* Este modelo contém as variáveis independentes a serem incluídas no modelo de regressão logística. Espera-se que o ajuste melhorará em relação ao modelo nulo e que resulte em um valor menor de $-2LL$.

Qualquer número de modelos propostos pode ser estimado (p.ex., modelos com uma, duas e três variáveis independentes podem ser propostas distintas).

3. *Avaliar a diferença –2LL*. O passo final é avaliar a significância estatística do valor –2*LL* entre os dois modelos (nulo versus proposto). Se os testes estatísticos suportam diferenças significantes, então podemos estabelecer que o conjunto de variáveis independentes no modelo proposto é significante na melhora do ajuste da estimação do mesmo.

De maneira semelhante, comparações também podem ser feitas entre dois modelos propostos quaisquer. Em tais casos, a diferença –2*LL* reflete a diferença em ajuste de modelo devido a distinções de especificações. Por exemplo, um modelo com duas variáveis independentes pode ser comparado com um modelo de três variáveis independentes para que se avalie a melhora pelo acréscimo de uma variável. Nesses casos, um modelo é escolhido para atuar como nulo e então é comparado com outro.

Por exemplo, considere que queremos testar a significância de um conjunto de variáveis independentes coletivamente para ver se elas melhoram o ajuste do modelo. O modelo nulo seria especificado como um modelo sem essas variáveis, e o modelo proposto incluiria as variáveis a serem avaliadas. A diferença em –2*LL* significaria a melhora a partir do conjunto de variáveis independentes. Poderíamos fazer testes similares das diferenças em –2*LL* entre outros pares de modelos variando o número de variáveis independentes incluídas em cada um.

O teste do qui-quadrado e o teste associado para significância estatística são usados para se avaliar a redução no logaritmo do valor de verossimilhança. No entanto, esses testes estatísticos são particularmente sensíveis a tamanho de amostra (para amostras pequenas é mais difícil mostrar significância estatística, e vice-versa para grandes amostras). Portanto, pesquisadores devem ser particularmente cuidadosos ao tirarem conclusões com base apenas na significância do teste do qui-quadrado em regressão logística.

Medidas pseudo R^2. Além dos testes qui-quadrado, diversas medidas do tipo R^2 foram desenvolvidas e são apresentadas em vários programas estatísticos para representarem ajuste geral do modelo. Essas medidas pseudo R^2 são interpretadas de uma maneira parecida com o coeficiente de determinação em regressão múltipla. Um valor **pseudo R^2** pode ser facilmente obtido para regressão logística semelhante ao valor R^2 em análise de regressão [6]. O pseudo R^2 para um modelo logit (R^2_{logit}) pode ser calculado como

$$R^2_{\text{LOGIT}} = \frac{-2LL_{\text{nulo}} - (-2LL_{\text{modelo}})}{-2LL_{\text{nulo}}}$$

Exatamente como na contraparte da regressão múltipla, o valor R^2 logit varia de 0,0 a 1,0. À medida que o modelo proposto aumenta o ajuste, o –2*LL* diminui. Um ajuste perfeito tem um valor de –2*LL* igual a 0,0 e um R^2_{LOGIT} de 1,0.

Duas outras medidas são semelhantes ao valor pseudo R^2 e são geralmente categorizadas também como medidas pseudo R^2. A medida R^2 de Cox e Snell opera do mesmo modo, com valores maiores indicando maior ajuste do modelo. No entanto, esta medida é limitada no sentido de que não pode atingir o valor máximo de 1, de forma que Nagelkerke propôs uma modificação que tinha o domínio de 0 a 1. Essas duas medidas adicionais são interpretadas como refletindo a quantia de variação explicada pelo modelo logístico, com 1,0 indicando ajuste perfeito.

Uma comparação com regressão múltipla. Ao discutir os procedimentos para avaliação de ajuste de modelo em regressão logística, fazemos várias referências a similaridades com regressão múltipla em termos de diversas medidas de ajuste. Na tabela a seguir, mostramos a correspondência entre conceitos usados em regressão múltipla e suas contrapartes em regressão logística.

Correspondência de elementos primários de ajuste de modelo

Regressão múltipla	Regressão logística
Soma total de quadrados	–2*LL* do modelo base
Soma de quadrados do erro	–2*LL* do modelo proposto
Soma de quadrados da regressão	Diferença de –*LL** para modelos base e proposto
Teste *F* de ajuste de modelo	Teste de qui-quadrado da diferença –2*LL*
Coeficiente de determinação (R^2)	Medidas pseudo R^2

Como podemos ver, os conceitos de regressão múltipla e regressão logística são semelhantes. Os métodos básicos para testar ajuste geral do modelo são comparáveis, com as diferenças surgindo dos métodos de estimação nas duas técnicas.

Precisão preditiva

Assim como emprestamos o conceito de R^2 da regressão como uma medida de ajuste geral de modelo, podemos procurar na análise discriminante a medida de precisão preditiva geral. As duas técnicas mais comuns são a matriz de classificação e as medidas de ajuste baseadas no qui-quadrado.

Matriz de classificação. Esta técnica de matriz de classificação é idêntica àquela usada em análise discriminante, ou seja, medir o quão bem a pertinência a grupo é prevista e desenvolver uma razão de sucesso. O caso da regressão

* N. de R. T.: A frase correta seria "Diferença de –2LL".

logística sempre incluirá somente dois grupos, mas todas as medidas relacionadas a chances (p.ex., chance máxima, chance proporcional ou *Q* de Press) usadas anteriormente são aplicáveis aqui também.

Medida baseada no qui-quadrado. Hosmer e Lemeshow [11] desenvolveram um teste de classificação no qual os casos são primeiramente divididos em aproximadamente 10 classes iguais. Em seguida, os números de eventos reais e previstos são comparados em cada classe com a estatística qui-quadrado. Esse teste fornece uma medida ampla de precisão preditiva que é baseada não no valor de verossimilhança, mas sim na real previsão da variável dependente. O uso apropriado desse teste requer um tamanho de amostra de pelo menos 50 casos para garantir que cada classe tenha pelo menos cinco observações e geralmente até mesmo uma amostra maior, uma vez que o número de eventos previstos nunca fica abaixo de 1. Além disso, a estatística qui-quadrado é sensível a tamanho da amostra, permitindo assim que essa medida encontre diferenças muito pequenas, estatisticamente significantes, quando o tamanho da amostra se torna grande.

Tipicamente examinamos tantas dessas medidas de ajuste de modelo quanto possível. Espera-se que uma convergência de indicações dessas medidas forneça o suporte necessário ao pesquisador para a avaliação do ajuste geral do modelo.

Teste da significância dos coeficientes

A regressão logística testa hipóteses sobre coeficientes individuais, como se faz na regressão múltipla. Em regressão múltipla, o teste estatístico era para ver se o coeficiente era significantemente diferente de 0. Um coeficiente nulo indica que o mesmo não tem impacto sobre a variável dependente. Em regressão logística, usamos também um teste estatístico para ver se o coeficiente logístico é diferente de 0. Lembre, contudo, que em regressão logística usando o logit como medida dependente, um valor de 0 corresponde à razão de desigualdade de 1,00 ou uma probabilidade de 0,50 – valores que indicam que a probabilidade é igual para cada grupo (i.e., novamente nenhum efeito da variável independente sobre a previsão de pertinência ao grupo).

Em regressão múltipla, o valor *t* é utilizado para avaliar a significância de cada coeficiente. Regressão logística usa uma estatística diferente, a **estatística Wald**. Ela provê a significância estatística para cada coeficiente estimado de forma que testes de hipóteses podem ocorrer exatamente como se faz em regressão múltipla. Se o coeficiente logístico é estatisticamente significante, podemos interpretá-lo em termos de como o mesmo impacta a probabilidade estimada e conseqüentemente a previsão de pertinência a grupo.

Interpretação dos coeficientes

Uma das vantagens da regressão logística é que precisamos saber apenas se um evento (compra ou não, risco de crédito ou não, falência de empresa ou sucesso) ocorreu ou não para definir um valor dicotômico como nossa variável dependente. No entanto, quando analisamos esses dados usando transformação logística, a regressão e seus coeficientes assumem um significado algo diferente daqueles encontrados na regressão com uma variável dependente métrica. Analogamente, cargas discriminantes de uma análise discriminante de dois grupos são interpretadas diferentemente a partir de um coeficiente logístico.

A partir do processo de estimação descrito anteriormente, sabemos que os coeficientes (B_0, B_1, B_2, …, B_n) são na verdade medidas das variações na proporção das probabilidades (as razões de desigualdades). No entanto, coeficientes logísticos são difíceis de interpretar em sua forma original, pois eles são expressos em termos de logaritmos quando usamos o logit como a medida dependente. Assim, a maioria dos programas de computador fornece também um **coeficiente logístico exponenciado**, que é apenas uma transformação (anti-logaritmo) do coeficiente logístico original. Desse modo, podemos usar os coeficientes logísticos originais ou exponenciados para a interpretação. Os dois tipos de coeficientes logísticos diferem no sentido da relação da variável independente com as duas formas da dependente, como mostrado aqui:

Coeficiente logístico	Reflete mudanças em...
Original	Logit (logaritmo da razão de desigualdades)
Exponenciado	Razão de desigualdades

Discutimos na próxima seção como cada forma do coeficiente reflete direção e magnitude da relação da variável independente, mas requer diferentes métodos de interpretação.

Direção da relação

A direção da relação (positiva ou negativa) reflete as mudanças na variável dependente associadas com mudanças na independente. Uma relação positiva significa que um aumento na variável independente é associado com um aumento na probabilidade prevista, e vice-versa para uma relação negativa. Veremos que a direção da relação é refletida diferentemente nos coeficientes logísticos original e exponenciado.

Interpretação da direção de coeficientes originais. O sinal dos coeficientes originais (positivo ou negativo) indica a direção da relação, como foi visto nos coeficientes de regressão. Um valor positivo aumenta a probabilidade, enquanto um negativo diminui a mesma, pois os coeficientes originais são expressos em termos de valores logit, onde um valor de 0,0 corresponde a um valor de razão de desigualdade de 1,0 e uma probabilidade de 0,50. Assim, números negativos são relativos a razões de desigualdades menores que 1,0 e probabilidades menores que 0,50.

Interpretação da direção de coeficientes exponenciados. Coeficientes exponenciados devem ser interpretados diferentemente, pois eles são os logaritmos dos coeficientes originais. Considerando o logaritmo, estamos na verdade estabelecendo o coeficiente exponenciado em termos de razões de desigualdades, o que significa que exponenciados não terão valores negativos. Como o logaritmo de 0 (sem efeito) é 1,0, um coeficiente exponenciado igual a 1,0 na verdade corresponde a uma relação sem direção. Assim, coeficientes exponenciados acima de 1,0 refletem uma relação positiva, e valores menores que 1,0 representam relações negativas.

Um exemplo de interpretação. Examinemos um exemplo simples para ver o que queremos dizer em termos de diferenças entre as duas formas de coeficientes logísticos.

> Se B_i (o coeficiente original) é positivo, sua transformação (exponencial do coeficiente) será maior que 1, o que significa que a razão de desigualdade aumentará para qualquer variação positiva da variável independente. Assim, o modelo tem uma maior probabilidade prevista de ocorrência. De modo semelhante, se B_i é negativo, o coeficiente exponenciado é menor que um e a razão de desigualdades diminui. Um coeficiente de zero se iguala a um valor de 1,0 no coeficiente exponenciado, o que resulta em nenhuma mudança na razão de desigualdades.

Uma discussão mais detalhada da interpretação de coeficientes, transformação logística e procedimentos de estimação pode ser encontrada em diversos textos [11].

Magnitude da relação

Para determinar quanto da probabilidade mudará dada uma variação de uma unidade na variável independente, o valor numérico do coeficiente deve ser avaliado. Exatamente como na regressão múltipla, os coeficientes para variáveis métricas e não-métricas devem ser interpretados de forma diferenciada, pois cada um reflete diferentes impactos sobre a variável dependente.

Interpretação da magnitude de variáveis independentes métricas. Para variáveis métricas, a questão é: quanto a probabilidade estimada varia por conta de uma variação unitária na variável independente? Em regressão múltipla, sabíamos que o coeficiente de regressão era o coeficiente angular da relação linear entre a medida independente e a dependente. Um coeficiente de 1,35 indicava que a variável dependente aumentava 1,35 unidades cada vez que a variável independente aumentava uma unidade. Em regressão logística, sabemos que temos uma relação não-linear limitada entre 0 e 1, e assim os coeficientes devem ser interpretados de forma diferente. Além disso, temos os dois coeficientes original e exponenciado para considerar.

Coeficientes logísticos originais. Apesar de mais apropriados para determinarem a direção da relação, os coeficientes logísticos originais são menos úteis na determinação da magnitude da relação. Eles refletem a variação no valor logit (logaritmo da razão de desigualdades), uma unidade de medida particularmente não compreensível na representação do quanto as probabilidades realmente variam.

Coeficientes logísticos exponenciados. Coeficientes exponenciados refletem diretamente a magnitude da variação no valor da razão de desigualdades. Por serem expoentes, eles são interpretados de maneira ligeiramente diferente. Seu impacto é multiplicativo, o que significa que o efeito do coeficiente não é adicionado à variável dependente (a razão de desigualdades), mas multiplicado para cada variação unitária na variável independente. Como tal, um coeficiente exponenciado de 1,0 denota mudança nenhuma (1,0 × variável independente = mudança nenhuma). Este resultado corresponde à nossa discussão anterior, onde coeficientes exponenciados menores que 1,0 refletem relações negativas, enquanto valores acima de 1,0 denotam relações positivas.

Um exemplo de avaliação da magnitude de variação. Talvez uma abordagem mais fácil para determinar a quantia de variação na probabilidade a partir desses valores seja como se segue:

$$\text{Mudança percentual na razão de desigualdades} = (\text{coeficiente exponenciado}_i - 1{,}0) \times 100$$

Os exemplos a seguir ilustram como calcular a variação de probabilidade devido a uma variação unitária na variável independente para um domínio de coeficientes exponenciados:

	Valor				
Coeficiente exponenciado (e^{b}_{i})	0,20	0,50	1,0	1,5	1,7
$e^{b}_{i} - 1{,}0$	−0,80	−0,50	0,0	0,50	0,70
Variação percentual na razão de desigualdades	−80%	−50%	0%	50%	70%

Se o coeficiente exponenciado é 0,20, uma mudança de uma unidade na variável independente reduzirá a razão de desigualdades em 80% (o mesmo se a razão de desigualdades fosse multiplicada por 0,20). Analogamente, um coeficiente exponenciado de 1,5 denota um aumento de 50% na razão de desigualdades.

Um pesquisador que conhece a razão de desigualdades existente e deseja calcular o novo valor dessa razão

para uma mudança na variável independente pode fazê-lo diretamente através do coeficiente exponenciado, como se segue:

Novo valor de razão de desigualdade = Valor antigo × Coeficiente exponenciado × Variação na variável independente

Usemos um exemplo simples para ilustrar a maneira como o coeficiente exponenciado afeta o valor da razão de desigualdades.

Considere que a razão de desigualdade é 1,0 (ou seja, 50-50) quando a variável independente tem um valor de 5,5 e o coeficiente exponenciado é 2,35. Sabemos que se este coeficiente for maior do que 1,0, então a relação é positiva, mas gostaríamos de saber o quanto a razão de desigualdades mudaria. Se esperamos que o valor da variável independente aumente 1,5 pontos para 7,0, podemos calcular o seguinte:

Nova razão de desigualdades = $1,0 \times 2,35 \times (7,0 - 5,5) = 3,525$

Razões de desigualdades podem ser traduzidas em termos de valores de probabilidade pela fórmula simples de Probabilidade = Razão de desigualdades/(1+Razão de desigualdades). Logo, a razão de 3,525 se traduz em uma probabilidade de 77,9% (3,25/(1 + 3,25)= 0,779), indicando que um aumento na variável independente de um ponto e meio aumenta a probabilidade de 50% para 78%, um aumento de 28%.

A natureza não-linear da curva logística é demonstrada, porém, quando novamente aplicamos o mesmo aumento à razão de desigualdades. Dessa vez, considere que a variável independente aumenta mais 1,5 pontos, para 8,5. Podemos esperar que a probabilidade aumente outros 28%? Não, pois isso faria a probabilidade ultrapassar os 100% (78% + 28% = 106%). Assim, o aumento ou diminuição da probabilidade diminui à medida que a curva se aproxima, mas jamais alcança, os dois pontos extremos (0 e 1). Neste exemplo, outro aumento de 1,5 cria um novo valor de razão de desigualdades de 12,426, traduzindo-se como uma razão de desigualdades de 92,6%, um aumento de 14%. Observe que neste caso de aumento de probabilidade a partir de 78%, o aumento na mesma para a variação de 1,5 na variável independente é metade (14%) daquilo que foi para o mesmo aumento quando a probabilidade era de 50%.

O pesquisador pode descobrir que coeficientes exponenciados são bastante úteis não apenas na avaliação do impacto de uma variável independente, mas no cálculo da magnitude dos efeitos.

Interpretação da magnitude para variáveis independentes não-métricas (dicotômicas). Como discutimos em regressão múltipla, variáveis dicotômicas representam uma única categoria de uma variável não-métrica (ver Capítulo 4 para uma discussão mais detalhada sobre o tema). Como tais, elas não são como variáveis métricas que variam em um intervalo de valores, mas assumem apenas os valores de 1 ou 0, indicando a presença ou ausência de uma característica. Como vimos na discussão anterior para variáveis métricas, os coeficientes exponenciados são a melhor maneira de interpretar o impacto da variável dicotômica, mas são interpretados diferentemente das variáveis métricas.

Sempre que uma variável dicotômica é usada, é essencial notar a categoria de referência ou omitida. Em uma maneira semelhante à interpretação em regressão, o coeficiente exponenciado representa o nível relativo da variável dependente para o grupo representado versus o grupo omitido. Podemos estabelecer essa relação como se segue:

$$\text{Razão de desigualdades}_{\substack{\text{categoria}\\\text{representada}}} = \text{Coeficiente exponenciado} \times \text{Razão de desigualdades}_{\substack{\text{categoria}\\\text{de referência}}}$$

Usemos um exemplo simples de dois grupos para ilustrar esses pontos.

Se a variável não-métrica é sexo, as duas possibilidades são masculino e feminino. A variável dicotômica pode ser definida como representando homens (i.e., valor 1 se for homem e 0 se for mulher) ou mulheres (i.e., valor 1 se for mulher e 0 se for homem). Qualquer que seja o caminho escolhido, porém, ele se determina como o coeficiente é interpretado. Consideremos que um valor 1 é dado às mulheres, fazendo com que o coeficiente exponenciado represente o percentual da razão de desigualdades de mulheres comparada com homens. Se o coeficiente é 1,25, então as mulheres têm uma razão de desigualdades 25% maior do que os homens (1,25 – 1,0 = 0,25). Analogamente, se o coeficiente é 0,80, então a razão de desigualdades para mulheres é 20% menor (0,80 – 1,0 = –0,20) do que para os homens.

Cálculo de probabilidades para um valor específico da variável independente

Na discussão anterior da distribuição assumida de possíveis variáveis dependentes, descrevemos uma curva em forma de S, ou logística. Para representar a relação entre as variáveis dependente e independentes, os coeficientes devem, na verdade, representar relações não-lineares entre as variáveis dependente e independentes. Apesar de o processo de transformação que envolve logaritmos fornecer uma linearização da relação, o pesquisador deve lembrar que os coeficientes na verdade correspondem a diferentes coeficientes angulares na relação ao longo dos valores da variável independente. Desse modo, a distribuição em forma de S pode ser estimada. Se o pesquisa-

dor estiver interessado no coeficiente angular da relação em vários valores da variável independente, os coeficientes podem ser calculados e a relação, avaliada [6].

Visão geral da interpretação dos coeficientes

A similaridade dos coeficientes com aqueles encontrados em regressão múltipla tem sido uma razão prioritária para a popularidade da regressão logística. Como vimos na discussão anterior, muitos aspectos são bastante semelhantes, mas o caráter único da variável dependente (a razão de desigualdades) e a forma logarítmica da variável estatística (necessitando uso dos coeficientes exponenciados) requer uma abordagem de algum modo de interpretação diferente. O pesquisador, contudo, ainda tem a habilidade para avaliar a direção e a magnitude do impacto de cada variável independente sobre a medida dependente e, em última instância, a precisão de classificação do modelo logístico.

Resumo

O pesquisador que se defronta com uma variável dependente dicotômica não precisa apelar para métodos elaborados para acomodar as limitações da regressão múltipla, e nem precisa ser forçado a empregar a análise discriminante, especialmente se suas suposições estatísticas são violadas. A regressão logística aborda esses problemas e fornece um método desenvolvido para lidar diretamente com essa situação da maneira mais eficiente possível.

UM EXEMPLO ILUSTRATIVO DE REGRESSÃO LOGÍSTICA

A regressão logística é uma alternativa atraente à análise discriminante sempre que a variável dependente tem apenas duas categorias. Suas vantagens em relação à análise discriminante incluem as seguintes:

1. É menos afetada do que a análise discriminante pelas desigualdades de variância-covariância ao longo dos grupos, uma suposição básica da análise discriminante.
2. Lida facilmente com variáveis independentes categóricas, enquanto na análise discriminante o uso de variáveis dicotômicas cria problemas com igualdades de variância-covariância.
3. Os resultados empíricos acompanham paralelamente os da regressão múltipla em termos de sua interpretação e das medidas diagnósticas de casos disponíveis para exame de resíduos.

O exemplo a seguir, idêntico ao da análise discriminante de dois grupos discutido anteriormente, ilustra essas vantagens e a similaridade da regressão logística com os resultados obtidos da regressão múltipla. Como veremos, ainda que a regressão logística tenha muitas vantagens como alternativa à análise discriminante, o pesquisador deve interpretar cuidadosamente os resultados devido aos aspectos ímpares de como a regressão logística lida com a previsão de probabilidades e de pertinência a grupos.

REGRAS PRÁTICAS 5-5

Regressão logística

- Regressão logística é o método preferido para variáveis dependentes de dois grupos (binárias) devido à sua robustez, facilidade de interpretação e diagnóstico
- Testes de significância de modelo são feitos com um teste de qui-quadrado sobre as diferenças no logaritmo da verossimilhança ($-2LL$) entre dois modelos
- Coeficientes são expressos em duas formas: original e exponenciado, para auxiliar na interpretação
- A interpretação dos coeficientes quanto a direção e magnitude é:
  - Direção pode ser avaliada diretamente nos coeficientes originais (sinais positivos ou negativos) ou indiretamente nos exponenciados (menor que 1 é negativa e maior que 1 é positiva)
  - Magnitude é avaliada melhor pelo coeficiente exponenciado, com a variação percentual na variável dependente mostrada por:

    Variação percentual = (Coeficiente exponenciado – 1,0) × 100

Estágios 1, 2 e 3: Objetivos da pesquisa, planejamento de pesquisa e suposições estatísticas

As questões abordadas nos primeiros três estágios do processo de decisão são idênticas para a análise discriminante de dois grupos e para a regressão logística.

O problema de pesquisa ainda é determinar se as diferenças de percepções de HBAT (X_6 a X_{18}) existem entre os clientes dos EUA/América do Norte e aqueles do resto do mundo (X_4). A amostra de 100 clientes é dividida em uma amostra de análise de 60 observações, com as 40 observações restantes constituindo a amostra de validação.

Agora nos concentramos sobre os resultados obtidos a partir do uso de regressão logística para estimar e compreender as diferenças entre esses dois tipos de clientes.

Estágio 4: Estimação do modelo de regressão logística e avaliação do ajuste geral

Antes que comece o processo de estimação, é possível rever as variáveis individuais e avaliar seus resultados univariados em termos de diferenciação entre grupos. Sabendo-se que os objetivos da análise discriminante e da regressão logística são os mesmos, podemos usar as mes-

mas medidas de discriminação para avaliar efeitos univariados, como foi feito para a análise discriminante.

Se revisarmos nossa discussão a respeito das diferenças dos grupos quanto às 13 variáveis independentes (olhar a Tabela 5-5), lembraremos que cinco variáveis (X_6, X_{11}, X_{12}, X_{13}, e X_{17}) tinham diferenças estatisticamente significantes entre os dois grupos. Se você olhar novamente a discussão no exemplo de dois grupos, lembre de uma indicação de multicolinearidade entre essas variáveis, pois ambas X_6 e X_{13} eram parte do fator Valor do produto derivado pela análise fatorial (ver Capítulo 3). A regressão logística é afetada por multicolinearidade entre as variáveis independentes de uma maneira semelhante à análise discriminante e análise de regressão.

Exatamente como em análise discriminante, essas cinco variáveis seriam as candidatas lógicas para inclusão na variável estatística de regressão logística, pois elas demonstram as maiores diferenças entre grupos. Regressão logística pode incluir uma ou mais dessas variáveis no modelo, bem como outras variáveis que não apresentam diferenças significantes neste estágio se elas operam em combinação com outras variáveis para significativamente melhorar a previsão.

Estimação do modelo

A regressão logística é estimada de maneira análoga à regressão múltipla, no sentido de que um modelo base é primeiramente estimado para fornecer um padrão para comparação (ver discussão anterior para maiores detalhes). Em regressão múltipla, a média é usada para estabelecer o modelo base e calcular a soma total de quadrados. Em regressão logística, o mesmo processo é empregado, com a média usada no modelo estimado não para estabelecer a soma de quadrados, mas para estabelecer o valor do logaritmo da verossimilhança. A partir desse modelo, podem ser estabelecidas as correlações parciais para cada variável e a variável mais discriminante pode ser escolhida de acordo com os critérios de seleção.

Estimação do modelo base. A Tabela 5-25 contém os resultados do modelo base para a análise de regressão logística. O valor do logaritmo da verossimilhança (–2*LL*) aqui é 82,108. A estatística escore, uma medida de associação usada em regressão logística, é a medida usada para selecionar variáveis no procedimento *stepwise*. Diversos critérios podem ser empregados para orientar a entrada: maior redução no valor –2*LL*, maior coeficiente de Wald, ou maior probabilidade condicional. Em nosso exemplo, empregamos o critério da redução da razão do logaritmo da verossimilhança.

Ao revermos a estatística de escores de variáveis não presentes no modelo neste momento, percebemos que as mesmas cinco variáveis com diferenças estatisticamente significantes (X_6, X_{11}, X_{12}, X_{13} e X_{17}) também são s únicas variáveis com estatística de escore significante na Tabela 5-25. Como o procedimento *stepwise* seleciona a variável com a maior estatística de escore, X_{13} deve ser a variável adicionada no primeiro passo.

Estimação *stepwise*: adição da primeira variável, X_{13}. Como esperado, X_{13} foi escolhida para entrada no primeiro passo do processo de estimação (ver Tabela

(*Continua*)

TABELA 5-25 Resultados do modelo base da regressão logística

Ajuste geral do modelo: medidas da qualidade do ajuste		
	Valor	
–2 Logaritmo de verossimilhança (–2*LL*)	82,108	
Variáveis fora da equação		
Variáveis independentes	**Estatística de escore**	**Significância**
X_6 Qualidade do produto	11,925	0,001
X_7 Atividades de comércio eletrônico	2,052	0,152
X_8 Suporte técnico	1,609	0,205
X_9 Solução de reclamação	0,866	0,352
X_{10} Anúncio	0,791	0,374
X_{11} Linha do produto	18,323	0,000
X_{12} Imagem da equipe de venda	8,622	0,003
X_{13} Preços competitivos	21,330	0,000
X_{14} Garantia e reclamações	0,465	0,495
X_{15} Novos produtos	0,614	0,433
X_{16} Encomenda e cobrança	0,090	0,764
X_{17} Flexibilidade de preço	21,204	0,000
X_{18} Velocidade de entrega	0,157	0,692

TABELA 5-26 Estimação *stepwise* da regressão logística: Adição de X_{13} (Preços competitivos)

Ajuste geral do modelo: medidas da qualidade de ajuste

		VARIAÇÃO EM –2*LL*			
		Do modelo base		***Do passo anterior***	
	Valor	**Variação**	**Significância**	**Variação**	**Significância**
–2 Logaritmo de verossimilhança (–2*LL*)	56,971	25,136	0,000	25,136	0,000
R^2 de Cox e Snell	0,342				
R^2 de Nagelkerke	0,459				
Pseudo R^2	0,306				

	Valor	**Significância**
χ^2 de Hosmer e Lemeshow	17,329	0,027

Variáveis na equação

Variável independente	**B**	**Erro padrão**	**Wald**	***df***	**Sig.**	**Exp(B)**
X_{13} Preços competitivos	1,129	0,287	15,471	1	0,000	3,092
Constante	–7,008	1,836	14,570	1	0,000	0,001

B = coeficiente logístico, Exp(B) = coeficiente exponenciado

Variáveis fora da equação

Variáveis independentes	**Estatística de escore**	**Significância**
X_6 Qualidade do produto	4,859	0,028
X_7 Atividades de comércio eletrônico	0,132	0,716
X_8 Suporte técnico	0,007	0,932
X_9 Solução de reclamação	1,379	0,240
X_{10} Anúncio	0,129	0,719
X_{11} Linha do produto	6,154	0,013
X_{12} Imagem da equipe de venda	2,745	0,098
X_{14} Garantia e reclamações	0,640	0,424
X_{15} Novos produtos	0,344	0,557
X_{16} Encomenda e cobrança	2,529	0,112
X_{17} Flexibilidade de preço	13,723	0,000
X_{18} Velocidade de entrega	1,206	0,272

Matriz de classificação

	Pertinência prevista em grupo[c]					
	AMOSTRA DE ANÁLISE[a]			AMOSTRA DE TESTE[b]		
	X_4 Região			***X_4 Região***		
Pertinência real em grupo	**EUA/América do Norte**	**Fora da América do Norte**	**Total**	**EUA/América do Norte**	**Fora da América do Norte**	**Total**
EUA/América do Norte	19 (73,1)	7	26	4 (30,8)	9	13
Fora da América do Norte	9	25 (73,5)	34	1	26 (96,3)	27

[a]73,3% de amostra de análise corretamente classificada.
[b]75,0% da amostra de teste corretamente classificada.
[c]Valores entre parênteses são percentuais corretamente classificados (razão de sucesso).

(*Continuação*)

5-26). Ela corresponde à maior estatística de escore em todas as 13 variáveis de percepções. A entrada de X_{13} no modelo de regressão logística conseguiu um razoável ajuste, com valores pseudo R^2 variando de 0,306 a 0,459 e as razões de sucesso de 73,3% e 75% para as amostras de análise e de teste, respectivamente.

O exame dos resultados, porém, identifica duas razões para se considerar um estágio extra para adicionar variáveis ao modelo de regressão logística:

- Três variáveis não presentes no modelo logístico corrente (X_{17}, X_{11} e X_6) têm estatísticas de escore estatisticamente significantes, indicando que a inclusão das mesmas melhoraria consideravelmente o ajuste geral do modelo.
- A razão de sucesso geral para a amostra de teste é boa (75,0%), mas um dos grupos (Clientes dos EUA/América do Norte) tem uma razão de sucesso inaceitavelmente baixa de 30,8%.

Estimação *stepwise*: Adição da segunda variável, X_{17}. Espera-se que um ou mais passos no procedimento *stepwise* resulte na inclusão de todas as variáveis independentes com estatística de escore significante, bem como sejam atingidas razões aceitáveis de sucesso (geral e específicas de grupos) tanto para a amostra de análise quanto para a de teste.

X_{17}, com a maior estatística de escore depois de adicionar X_{13}, foi escolhida para entrada no passo 2 (Tabela 5-27). Melhoras em todas as medidas de ajuste de modelo variaram de uma queda no valor $-2LL$ até as várias medidas R^2. Mais importante sob uma perspectiva de estimação de modelo, nenhuma das variáveis fora da equação tinha variações estatisticamente significantes de escores.

Assim, o modelo logístico de duas variáveis incluindo X_{13} e X_{17} será o modelo final a ser usado para fins de avaliação de ajuste do mesmo, de precisão preditiva e de interpretação dos coeficientes.

Avaliação do ajuste geral do modelo

Ao se fazer uma avaliação do ajuste geral de um modelo logístico de regressão, podemos empregar três abordagens: medidas estatísticas de ajuste geral do modelo, medidas pseudo R^2, e precisão de classificação expressada na razão de sucesso. Cada uma dessas abordagens será examinada para os modelos de regressão logística de uma variável e de duas variáveis que resultaram do procedimento *stepwise*.

Medidas estatísticas. A primeira medida estatística é o teste qui-quadrado para a variação no valor $-2LL$ do modelo base, que é comparável com o teste F geral em regressão múltipla. Valores menores da medida $-2LL$ indicam um melhor ajuste de modelo, e o teste estatístico está disponível para avaliar a diferença entre o modelo base e os demais modelos propostos (em um procedimento *stepwise*, este teste está sempre baseado na melhora do passo anterior).

- No modelo de uma só variável (ver Tabela 5-26), o valor $-2LL$ é reduzido a partir do valor do modelo base de 82,108 para 59,971*, uma queda de 25,136. Este aumento em ajuste de modelo foi estatisticamente significante no nível 0,000.
- No modelo de duas variáveis, o valor $-2LL$ diminuiu mais para 39,960, resultando em quedas significantes não apenas do modelo base (42,148), mas também uma queda significante do modelo de uma variável (17,011). Ambas as melhoras de ajuste foram significantes no nível 0,000.

A segunda medida estatística é a de Hosmer e Lemeshow de ajuste geral [11]. Este teste estatístico mede a correspondência dos valores reais e previstos da variável dependente. Neste caso, um ajuste melhor de modelo é indicado por uma diferença menor na classificação observada e prevista.

O teste de Hosmer e Lemeshow mostra significância para o modelo logístico de uma variável (0,027 da Tabela 5-26), indicando que diferenças significantes ainda permanecem entre valores reais e esperados. O modelo de duas variáveis, contudo, reduz o nível de significância para 0,722 (ver Tabela 5-27), um valor não-significante que aponta para um ajuste aceitável.

Para o modelo logístico de duas variáveis, ambas as medidas estatísticas de ajuste geral do modelo indicam que o mesmo é aceitável e em um nível estatisticamente significante. No entanto, é necessário examinar as outras medidas de ajuste geral do modelo para avaliar se os resultados alcançam os níveis necessários de significância prática também.

Medidas de pseudo R^2. Três medidas disponíveis são comparáveis com a medida R^2 em regressão múltipla: R^2 de Cox e Snell, R^2 de Nagelkerke, e a medida pseudo R^2 baseada na redução no valor $-2LL$.

Para o modelo de regressão logística de uma variável, esses valores eram 0,342, 0,459 e 0,306, respectivamente. Combinados, eles indicam que o modelo de regressão de uma variável explica aproximadamente um terço da variação na medida dependente. Apesar de o modelo de uma variável ser considerado estatisticamente significante em diversas medidas de ajuste geral, esses valores de R^2 são um pouco baixos para fins de significância prática.

(*Continua*)

* N. de R. T.: O número correto é 56,971.

TABELA 5-27 Estimação *stepwise* da regressão logística: adição de X_{17} (Flexibilidade de preços)

Ajuste geral do modelo: medidas da qualidade de ajuste

		VARIAÇÃO EM –2*LL*			
		Do modelo base		***Do passo anterior***	
	Valor	**Variação**	**Significância**	**Variação**	**Significância**
–2 Logaritmo de verossimilhança (–2*LL*)	39,960	42,148	0,000	17,011	0,000
R^2 de Cox e Snell	0,505				
R^2 de Nagelkerke	0,677				
Pseudo R^2	0,513				

	Valor	**Significância**
χ^2 de Hosmer e Lemeshow	5,326	0,722

Variáveis na equação

Variável independente	**B**	**Erro padrão**	**Wald**	***df***	**Sig.**	**Exp(B)**
X_{13} Preços competitivos	1,079	0,357	9,115	1	0,003	2,942
X_{17} Flexibilidade de preços	1,844	0,639	8,331	1	0,004	6,321
Constante	–14,192	3,712	14,614	1	0,000	0,000

B = coeficiente logístico, Exp(B) = coeficiente exponenciado

Variáveis fora da equação

Variáveis independentes	**Estatística de escore**	**Significância**
X_6 Qualidade do produto	0,656	0,418
X_7 Atividades de comércio eletrônico	3,501	0,061
X_8 Suporte técnico	0,006	0,937
X_9 Solução de reclamação	0,693	0,405
X_{10} Anúncio	0,091	0,762
X_{11} Linha do produto	3,409	0,065
X_{12} Imagem da equipe de venda	0,849	0,357
X_{14} Garantia e reclamações	2,327	0,127
X_{15} Novos produtos	0,026	0,873
X_{16} Encomenda e cobrança	0,010	0,919
X_{18} Velocidade de entrega	2,907	0,088

Matriz de classificação

	Pertinência prevista em grupo[c]					
	AMOSTRA DE ANÁLISE[a]			AMOSTRA DE TESTE[b]		
	X_4 Região			***X_4 Região***		
Pertinência real em grupo	**EUA/América do Norte**	**Fora da América do Norte**	**Total**	**EUA/América do Norte**	**Fora da América do Norte**	**Total**
EUA/América do Norte	25	1	26	9	4	13
	(96,2)			(69,2)		
Fora da América do Norte	6	28	34	2	25	27
		(82,4)			(92,6)	

[a]88,3% de amostra de análise corretamente classificada.
[b]85,0% da amostra de teste corretamente classificada.
[c]Valores entre parênteses são percentuais corretamente classificados (razão de sucesso).

(*Continuação*)

O modelo de duas variáveis (ver Tabela 5-27) tem valores R^2 que são ambos maiores que 0,50, apontando para um modelo de regressão logística que explica pelo menos metade da variação entre os dois grupos de clientes. Sempre se deseja melhorar tais valores, mas tal nível é considerado praticamente significante nesta situação.

Os valores R^2 do modelo de duas variáveis exibiram considerável melhora sobre o modelo de uma variável e indicam bom ajuste quando comparados aos valores R^2 geralmente encontrados em regressão múltipla. De acordo com as medidas de ajuste de caráter estatístico, o modelo é considerado aceitável em termos de significância estatística e prática.

Precisão de classificação. O terceiro exame de ajuste geral do modelo será para avaliar a precisão de classificação do modelo em uma medida final de significância prática. As matrizes de classificação, idênticas em natureza àquelas empregadas em análise discriminante, representam os níveis de precisão preditiva atingidos pelo modelo logístico. A medida de precisão preditiva usada é a razão de sucesso, o percentual de casos corretamente classificados. Esses valores serão calculados tanto para a amostra de análise quanto a de teste, e medidas específicas de grupos serão examinadas além das medidas gerais. Além disso, comparações podem ser feitas, como ocorreu em análise discriminante, com padrões de comparação representando os níveis de precisão preditiva conseguidos por chances (ver discussão mais detalhada na seção sobre análise discriminante).

Os padrões de comparação para as razões de sucesso da matriz de classificação serão os mesmos que foram calculados para a análise discriminante de dois grupos. Os valores são 65,5% para o critério de chance proporcional (a medida preferida) e 76,3% para o critério de chance máxima. Se você não estiver familiarizado com os métodos de cálculo de tais medidas, veja a discussão anterior no capítulo que trata de avaliação da precisão de classificação.

- As razões de sucesso geral para o modelo logístico de uma variável são 73,3% e 75,0% para as amostras de análise e de teste, respectivamente. Mesmo que as razões de sucesso geral sejam maiores do que o critério de chance proporcional e comparáveis com o critério de chance máxima, um problema considerável surge na amostra de teste para os clientes dos EUA/América do Norte, onde a razão de sucesso é de somente 30,8%. Este nível está abaixo de ambos os padrões e demanda que o modelo logístico seja expandido até o ponto em que, espera-se, esta razão de sucesso específica de grupo exceda os padrões.
- O modelo de duas variáveis exibe melhora substancial na razão de sucesso geral e nos valores específicos de grupos. As razões de sucesso geral subiram para 88,3% e 85,0% para as amostras de análise e de teste, respectivamente. Além disso, a problemática razão de sucesso na amostra de teste aumenta para 69,2%, acima do valor padrão para o critério de chance proporcional. Com essas melhoras nos níveis geral e específicos, o modelo de regressão logística de duas variáveis é considerado aceitável em termos de precisão de classificação.

Em todos os três dos tipos básicos de medida de ajuste geral, o modelo de duas variáveis (com X_{13} e X_{17}) demonstra níveis aceitáveis de significância estatística e prática. Com ajuste de modelo geral aceitável, voltamos nossa atenção para a avaliação dos testes estatísticos dos coeficientes logísticos a fim de identificar os coeficientes que têm relações significantes afetando pertinência a grupo.

Significância estatística dos coeficientes

Os coeficientes estimados para as duas variáveis independentes e a constante também podem ser avaliados quanto à significância estatística. A estatística Wald é usada para avaliar significância de um modo semelhante ao teste *t* utilizado em regressão múltipla.

Os coeficientes logísticos para X_{13} (1,079) e X_{17} (1,844) e a constante (–14,190*) são todos significantes no nível 0,01 com base no teste estatístico de Wald. Nenhuma outra variável consegue entrar no modelo e atingir pelo menos um nível de significância de 0,05.

Assim, as variáveis individuais são significantes e podem ser interpretadas para identificar as relações que afetam as probabilidades previstas e subseqüentemente a pertinência a grupo.

Diagnósticos por casos

A análise da má classificação de observações individuais pode fornecer uma melhor visão sobre possíveis melhoramentos do modelo. Diagnósticos por casos, como resíduos e medidas de influências, estão disponíveis, bem como a análise de perfil discutida anteriormente para a análise discriminante.

Neste caso, apenas 13 casos foram mal classificados (7 na amostra de análise e 6 na de teste). Dado o elevado grau de correspondência entre esses casos e aqueles mal classificados estudados na análise discriminante de dois grupos, o processo de estabelecimento de perfil não será novamente levado adiante (leitores interessados podem rever o exemplo de dois grupos). Diagnóstico por casos,

* N. de R. T.: O número correto é -14,192.

como resíduos e medidas de influência estão disponíveis. Dados os baixos níveis de má classificação, porém, nenhuma análise complementar de classificação ruim é executada.

Estágio 5: Interpretação dos resultados

O procedimento de regressão logística *stepwise* produziu uma variável estatística muito semelhante àquela da análise discriminante de dois grupos, apesar de ter uma variável independente a menos. Examinaremos os coeficientes logísticos para avaliarmos a direção e o impacto que cada variável tem sobre a probabilidade prevista e a pertinência a grupo.

Interpretação dos coeficientes logísticos

O modelo final de regressão logística inclui duas variáveis (X_{13} e X_{17}) com coeficientes de regressão de 1,079 e 1,844, respectivamente, e uma constante de –14,190* (ver Tabela 5-27). A comparação desses resultados com a análise discriminante de dois grupos revela resultados quase idênticos, uma vez que a análise discriminante incluiu três variáveis no modelo de dois grupos – X_{13} e X_{17} juntamente com X_{11}.

Direção das relações. Para avaliar a direção da relação de cada variável, podemos examinar ou os coeficientes logísticos originais, ou os coeficientes exponenciados. Comecemos com os originais.

Se você recordar de nossa discussão anterior, podemos interpretar a direção da relação diretamente a partir do sinal dos coeficientes logísticos originais. Neste caso, ambas as variáveis têm sinais positivos, o que aponta para uma relação positiva entre ambas as variáveis independentes e a probabilidade prevista. À medida que os valores de X_{13} ou X_{17} aumentam, a probabilidade prevista aumenta, fazendo crescer assim a possibilidade de que um cliente seja categorizado como residindo fora da América do Norte.

Voltando nossa atenção para os coeficientes expoenciados, devemos recordar que valores acima de 1,0 indicam uma relação positiva e valores abaixo de 1,0 apontam para uma relação negativa. Em nosso caso, os valores de 2,942 e 6,319 também refletem relações positivas.

Magnitude das relações. O método mais direto para avaliar a magnitude da variação na probabilidade devido a cada variável independente é examinar os coeficientes exponenciados. Como você deve lembrar, o coeficiente exponenciado menos um é igual à variação percentual da razão de desigualdades.

Em nosso caso, isso significa que um aumento de um ponto aumenta a razão de desigualdades em 194% para X_{13} e 531% para X_{17}. Esses números podem exceder 100% de variação porque eles estão aumentando a razão de desigualdades e não as probabilidades propriamente ditas. Os impactos são grandes porque o termo constante (–14,190*) define um ponto inicial de quase zero para os valores de probabilidade. Logo, grandes aumentos na razão de desigualdades são necessários para se conseguir valores maiores de probabilidades.

Outra abordagem na compreensão sobre como os coeficientes logísticos definem probabilidade é calcular a probabilidade prevista para qualquer conjunto de valores para as variáveis independentes.

Para as variáveis independentes X_{13} e X_{17}, usemos as médias para os dois grupos. Dessa maneira podemos ver qual seria a probabilidade prevista para um membro médio de cada grupo.

A Tabela 5-28 mostra os cálculos para a previsão da probabilidade para os dois centróides de grupo. Como podemos perceber, o centróide para o grupo 0 (clientes na América do Norte) tem uma probabilidade prevista de 18,9%, enquanto o centróide para o grupo 1 (fora da América do Norte) tem uma probabilidade prevista de 94,8%. Este exemplo demonstra que o modelo logístico cria de fato uma separação entre os dois centróides de grupo em termos de probabilidade prevista, gerando excelentes resultados de classificação conquistados para as amostras de análise e de teste.

Os coeficientes logísticos definem relações positivas para ambas as variáveis independentes e fornecem uma maneira de avaliar o impacto de uma variação em uma ou ambas as variáveis sobre a razão de desigualdades e conseqüentemente sobre a probabilidade prevista. Fica evidente por que muitos pesquisadores preferem regressão logística à análise discriminante quando comparações são feitas sobre a informação mais útil disponível nos coeficientes logísticos em contrapartida com as cargas discriminantes.

Estágio 6: Validação dos resultados

A validação do modelo de regressão logística é conseguida neste exemplo através do mesmo método usado em análise discriminante: criação de amostras de análises e de teste. Examinando a razão de sucesso para a amostra de teste, o pesquisador pode avaliar a validade externa e a significância prática do modelo de regressão logística.

Para o modelo final de regressão logística de duas variáveis, as razões de sucesso para as amostras de análi-

(*Continua*)

* N. de R. T.: O número correto é -14,192.

TABELA 5-28 Cálculo de valores de probabilidade estimada para os centróides de grupos da região X_4

	X_4 (Região)	
	Grupo 0: EUA/América do Norte	**Grupo 1: Fora da América do Norte**
Centróide: X_{13}	5,60	7,42
Centróide: X_{17}	3,63	4,93
Valor logit[a]	–1,452	2,909
Razão de desigualdades[b]	0,234	18,332
Probabilidade[c]	0,189	0,948

[a]Calculado como: Logit = $-14{,}190 + 1{,}079X_{13} + 1{,}844X_{17}$
[b]Calculada como: Razão de desigualdades = e^{Logit}
[c]Calculada como: Probabilidade = Razão de desigualdades/(1+Razão de desigualdades)

(*Continuação*)
se e de teste excedem todos os padrões de comparação (critérios de chance proporcional e de chance máxima). Além disso, todas as razões de sucesso específicas de grupos são suficientemente grandes para a aceitação. Esse aspecto é especialmente importante para a amostra de teste, que é o principal indicador de validade externa.

Esses resultados levam à conclusão de que o modelo de regressão logística, como também descoberto com o modelo de análise discriminante, demonstrou validade externa suficiente para a completa aceitação dos resultados.

Uma visão gerencial

A regressão logística apresenta uma alternativa à analise discriminante que pode ser mais confortável para muitos pesquisadores devido à sua similaridade com regressão múltipla. Dada a sua robustez diante das condições de dados que podem afetar negativamente a análise discriminante (p.ex., matrizes diferentes de variância-covariância), a regressão logística é também a técnica preferida de estimação em muitas aplicações.

Quando comparada com análise discriminante, a regressão logística fornece precisão preditiva comparável com uma variável estatística mais simples que usava a mesma interpretação substancial, apenas com uma variável a menos. A partir dos resultados da regressão logística, o pesquisador pode se concentrar na competitividade e na flexibilidade de preços como as principais variáveis de diferenciação entre os dois grupos de clientes. A meta nesta análise não é aumentar probabilidade (como poderia ser o caso de se analisar sucesso versus fracasso), ainda que a regressão logística forneça uma técnica direta para a HBAT compreender o impacto relativo de cada variável independente na criação de diferenças entre os dois grupos de clientes.

Resumo

A natureza intrínseca, os conceitos e a abordagem para a análise discriminante múltipla e a regressão logística foram apresentadas. Orientações básicas para sua aplicação e interpretação foram incluídas para melhor esclarecer os conceitos metodológicos. Este capítulo ajuda você a fazer o seguinte:

Estabelecer as circunstâncias sob as quais a análise discriminante linear ou a regressão logística devem ser usadas ao invés da regressão múltipla. Ao se escolher uma técnica analítica apropriada, às vezes encontramos um problema que envolve uma variável dependente categórica e diversas variáveis independentes métricas. Lembre-se que a variável dependente em regressão foi medida metricamente. Análise discriminante múltipla e regressão logística são as técnicas estatísticas apropriadas quando o problema de pesquisa envolve uma única variável dependente categórica e diversas variáveis independentes métricas. Em muitos casos, a variável dependente consiste de dois grupos ou classificações, por exemplo, masculino versus feminino, alto versus baixo, ou bom versus ruim. Em outros casos, mais de dois grupos estão envolvidos, como classificações baixas, médias e altas. A análise discriminante e a regressão logística são capazes de lidar com dois ou múltiplos (três ou mais) grupos. Os resultados de uma análise discriminante e de uma regressão logística podem auxiliar no perfil das características entre-grupos dos indivíduos e na correspondência dos mesmos com seus grupos adequados.

Identificar os principais problemas relacionados aos tipos de variáveis usados e os tamanhos de amostras exigidos na aplicação de análise discriminante. Para aplicar análise discriminante, o pesquisador deve primeiramente especificar quais variáveis devem ser medidas independentes e qual é a dependente. O pesquisador deve se concentrar primeiro na variável dependente. O número de grupos da variável dependente (categorias) pode ser dois ou mais, mas tais grupos devem ser mutuamente excludentes e exaustivos. Depois que uma decisão foi tomada sobre a variável dependente, o pesquisador deve decidir quais

variáveis independentes devem ser incluídas na análise. Variáveis independentes são escolhidas de duas maneiras: (1) identificando variáveis de pesquisa anterior ou do modelo teórico inerente à questão de pesquisa, e (2) utilizando o conhecimento e a intuição do pesquisador para selecionar variáveis para as quais nenhuma pesquisa ou teoria anterior existem mas que logicamente podem estar relacionadas com a previsão de grupos da variável dependente.

A análise discriminante, como as demais técnicas multivariadas, é afetada pelo tamanho da amostra sob análise. Uma proporção de 20 observações para cada variável preditora é recomendada. Como os resultados se tornam instáveis à medida que o tamanho da amostra diminui relativamente ao número de variáveis independentes, o tamanho mínimo recomendado é de cinco observações por variável independente. O tamanho amostral de cada grupo também deve ser considerado. No mínimo, o tamanho do menor grupo de uma categoria deve exceder o número de variáveis independentes. Como orientação prática, cada categoria deve ter pelo menos 20 observações. Mesmo que todas as categorias ultrapassem 20 observações, porém, o pesquisador também deve considerar os tamanhos relativos dos grupos. Variações grandes nos tamanhos dos grupos afetam a estimação da função discriminante e a classificação de observações.

Compreender as suposições subjacentes à análise discriminante na avaliação de sua adequação a um problema em particular. As suposições da análise discriminante se relacionam aos processos estatísticos envolvidos nos procedimentos de estimação e classificação, bem como aos problemas que afetam a interpretação dos resultados. As suposições-chave para se obter a função discriminante são normalidade multivariada das variáveis independentes, e estruturas (matrizes) desconhecidas (mas iguais) de dispersão e covariância para os grupos como definidos pela variável dependente. Se as suposições são violadas, o pesquisador deve entender o impacto sobre os resultados que podem ser esperados e considerar métodos alternativos para análise (p.ex., regressão logística).

Descrever as duas abordagens computacionais para análise discriminante e o método para avaliação de ajuste geral do modelo. As duas técnicas para análise discriminante são os métodos simultâneo (direto) e *stepwise*. A estimação simultânea envolve a computação da função discriminante considerando todas as variáveis independentes ao mesmo tempo. Portanto, a função discriminante é computada com base no conjunto inteiro de variáveis independentes, independentemente do poder discriminante de cada variável independente. A estimação *stepwise* é uma alternativa ao método simultâneo. Ela envolve a entrada de variáveis independentes uma por vez com base no poder discriminante das mesmas. O método *stepwise* segue um processo seqüencial de adição ou eliminação de variáveis da função discriminante. Depois que esta é estimada, o pesquisador deve avaliar a significância ou ajuste da mesma. Quando um método simultâneo é empregado, o lambda de Wilks, o traço de Hotelling e o critério de Pillai calculam a significância estatística do poder discriminatório da função estimada. Se um método *stepwise* é usado para estimar a função discriminante, o D^2 de Mahalanobis e a medida V de Rao são os mais adequados para avaliar ajuste.

Explicar o que é uma matriz de classificação e como desenvolver uma, e descrever as maneiras de se avaliar a precisão preditiva da função discriminante. Os testes estatísticos para avaliar a significância das funções discriminantes avaliam apenas o grau de diferença entre grupos com base nos escores Z discriminantes, mas não indicam o quão bem as funções prevêem. Para determinar a habilidade preditiva de uma função discriminante, o pesquisador deve construir matrizes de classificação. O procedimento da matriz de classificação fornece uma perspectiva sobre significância prática no lugar de significância estatística. Antes que uma matriz de classificação possa ser construída, no entanto, o pesquisador deve determinar o escore de corte para cada função discriminante. O escore de corte representa o ponto de divisão utilizado para classificar observações em cada um dos grupos, baseado no escore da função discriminante. O cálculo de um escore de corte entre dois grupos quaisquer é sustentado pelos dois centróides de grupo (média dos escores discriminantes) e pelos tamanhos relativos dos dois grupos. Os resultados do procedimento de classificação são apresentados em forma matricial. As entradas na diagonal da matriz representam o número de indivíduos corretamente classificados. Os números fora da diagonal correspondem a classificações incorretas. O percentual corretamente classificado, também conhecido como *razão de sucesso*, revela o quão bem a função discriminante prevê os objetos. Se os custos da má classificação forem aproximadamente iguais para todos os grupos, o escore de corte ótimo será aquele que classificar mal o menor número de objetos ao longo de todos os grupos. Se os custos de má classificação forem desiguais, o escore de corte ótimo será aquele que minimiza os custos de má classificação. Para avaliar a razão de sucesso, devemos olhar para uma classificação por chances. Quando os tamanhos de grupos são iguais, a determinação da classificação por chances se baseia no número de grupos. Quando os tamanhos dos grupos são distintos, o cálculo da classificação por chances pode ser feito de duas maneiras: chance máxima e chance proporcional.

Dizer como identificar variáveis independentes com poder discriminatório. Se a função discriminante é estatisticamente significante e a precisão de classificação (razão de sucesso) é aceitável, o pesquisador deve se concentrar na realização de interpretações substanciais das descobertas. Este processo envolve a determinação da importância

relativa de cada variável independente na discriminação entre os grupos. Três métodos de determinação da importância relativa foram propostos: (1) pesos discriminantes padronizados, (2) cargas discriminantes (correlações estruturais) e (3) valores *F* parciais. A abordagem tradicional para interpretar funções discriminantes examina o sinal e a magnitude do peso discriminante padronizado designado para cada variável na computação das funções discriminantes. Variáveis independentes com pesos relativamente maiores contribuem mais para o poder discriminatório da função do que variáveis com pesos menores. O sinal denota se a variável contribui negativa ou positivamente. Cargas discriminantes são cada vez mais usadas como uma base para interpretação por conta das deficiências na utilização de pesos. Medindo a correlação linear simples entre cada variável independente e a função discriminante, as cargas discriminantes refletem a variância que as variáveis independentes compartilham com a função discriminante. Elas podem ser interpretadas como cargas fatoriais na avaliação da contribuição relativa de cada variável independente à função discriminante. Quando um método de estimação *stepwise* é usado, uma maneira adicional de interpretar o poder discriminatório relativo das variáveis independentes é através do emprego de valores *F* parciais, o que se consegue examinando-se os tamanhos absolutos dos valores *F* significantes e ordenando-os. Valores *F* grandes indicam um poder discriminatório maior.

Justificar o uso de um método de divisão de amostra para validação. O estágio final de uma análise discriminante envolve a validação dos resultados discriminantes para fornecer garantias de que os mesmos têm tanto validade interna quanto externa. Além de validar as razões de sucesso, o pesquisador deve usar o perfil dos grupos para garantir que as médias deles são indicadores válidos do modelo conceitual utilizado na seleção das variáveis independentes. Validação pode ocorrer com uma amostra separada (de teste) ou utilizando um procedimento que repetidamente processa a amostra de estimação. Validação das razões de sucesso é executada muito freqüentemente criando-se uma amostra de teste, também chamada de amostra de validação. O propósito da utilização de uma amostra de teste para fins de validação é perceber o quão bem a função discriminante funciona em uma amostra de observações que não foram usadas para obtê-la. Tal avaliação envolve o desenvolvimento de uma função discriminante com a amostra de análise e então a aplicação da função à amostra de teste.

Entender as vantagens e desvantagens da regressão logística comparada com análise discriminante e regressão múltipla. Análise discriminante é apropriada quando a variável dependente é não-métrica. Se ela tiver apenas dois grupos, então a regressão logística pode ser preferível por duas razões. Primeiro, a análise discriminante apóia-se no atendimento estrito das suposições de normalidade multivariada e igualdade entre as matrizes de variância-covariância nos grupos – premissas que não são atendidas em muitas situações. A regressão logística não se depara com tais restrições e é muito mais robusta quando essas suposições não são atendidas, tornando sua aplicação adequada em muitos casos. Segundo, mesmo que as suposições sejam atendidas, muitos pesquisadores preferem a regressão logística por ser semelhante à regressão múltipla. Como tal, ela tem testes estatísticos diretos, métodos semelhantes para incorporar variáveis métricas e não-métricas e efeitos não-lineares, bem como uma vasta gama de diagnósticos. A regressão logística é equivalente à análise discriminante de dois grupos e pode ser mais adequada em muitas situações.

Interpretar os resultados de uma análise de regressão logística, com comparações com regressão múltipla e análise discriminante. A adequação de ajuste para um modelo de regressão logística pode ser avaliada de duas maneiras: (1) usando valores pseudo R^2, semelhantes àqueles encontrados em regressão múltipla, e (2) examinando precisão preditiva (i.e., a matriz de classificação em análise discriminante). As duas abordagens examinam ajuste de modelo sob diferentes perspectivas, mas devem conduzir a resultados semelhantes. Uma das vantagens da regressão logística é que precisamos saber apenas se um evento ocorreu para definir um valor dicotômico como nossa variável dependente. Quando analisamos esses dados usando transformação logística, contudo, a regressão logística e seus coeficientes assumem um significado um tanto diferente daqueles encontrados em regressão com uma variável dependente métrica. Analogamente, cargas em análise discriminante são interpretadas diferentemente de um coeficiente logístico. Este último reflete a direção e a magnitude da relação da variável independente, mas requer diferentes métodos de interpretação. A direção da relação (positiva ou negativa) retrata as variações na variável dependente associadas com mudanças na independente. Uma relação positiva significa que um aumento na variável independente é associado com um aumento na probabilidade prevista, e vice-versa para uma relação negativa.Para determinar a magnitude do coeficiente, ou o quanto que a probabilidade mudará dada uma unidade de variação na variável independente, o valor numérico do coeficiente deve ser avaliado. Exatamente como em regressão múltipla, os coeficientes para variáveis métricas e não-métricas devem ser interpretados diferentemente porque cada um reflete diferentes impactos sobre a variável dependente.

A análise discriminante múltipla e a regressão logística ajudam a compreender e explicar problemas de pesquisa que envolvem uma variável dependente categórica e diversas variáveis independentes métricas. Ambas as técnicas podem ser usadas para estabelecer o perfil das

características entre grupos dos indivíduos e designar os mesmos a seus grupos apropriados. Aplicações potenciais dessas duas técnicas tanto em negócios como em outras áreas são inúmeras.

Questões

1. Como você diferenciaria entre análise discriminante múltipla, análise de regressão, regressão logística e análise de variância?
2. Quando você empregaria regressão logística no lugar de análise discriminante? Quais são as vantagens e desvantagens dessa decisão?
3. Quais critérios você poderia usar para decidir se deve parar uma análise discriminante depois de estimar a função discriminante? Depois do estágio de interpretação?
4. Qual procedimento você seguiria para dividir sua amostra em grupos de análise e de teste? Como você mudaria este procedimento se sua amostra consistisse de menos do que 100 indivíduos ou objetos?
5. Como você determinaria o escore de corte ótimo?
6. Como você determinaria se a precisão de classificação da função discriminante é suficientemente alta relativamente a uma classificação ao acaso?
7. Como uma análise discriminante de dois grupos difere de uma análise de três grupos?
8. Por que um pesquisador deve expandir as cargas e dados do centróide ao representar graficamente uma solução de análise discriminante?
9. Como a regressão logística e a análise discriminante lidam com a relação das variáveis dependente e independentes?
10. Quais são as diferenças de estimação e interpretação entre regressão logística e análise discriminante?
11. Explique o conceito de razão de desigualdades e por que ela é usada para prever probabilidade em um procedimento de regressão logística.

Leituras sugeridas

Uma lista de leituras sugeridas ilustrando questões e aplicações da análise discriminante e regressão logística está disponível na Web em www.prenhall.com/hair (em inglês).

Referências

1. Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.
2. Crask, M., and W. Perreault. 1977. Validation of Discriminant Analysis in Marketing Research. *Journal of Marketing Research* 14 (February): 60–68.
3. Demaris, A. 1995. A Tutorial in Logistic Regression. *Journal of Marriage and the Family* 57: 956–68.
4. Dillon, W. R., and M. Goldstein. 1984. *Multivariate Analysis: Methods and Applications.* New York: Wiley.
5. Frank, R. E., W. E. Massey, and D. G. Morrison. 1965. Bias in Multiple Discriminant Analysis. *Journal of Marketing Research* 2(3): 250–58.
6. Gessner, Guy, N. K. Maholtra, W. A. Kamakura, and M. E. Zmijewski. 1988. Estimating Models with Binary Dependent Variables: Some Theoretical and Empirical Observations. *Journal of Business Research* 16(1): 49–65.
7. Green, P. E., D. Tull, and G. Albaum. 1988. *Research for Marketing Decisions.* Upper Saddle River, NJ: Prentice Hall.
8. Green, P. E. 1978. *Analyzing Multivariate Data.* Hinsdale, IL: Holt, Rinehart and Winston.
9. Green, P. E., and J. D. Carroll. 1978. *Mathematical Tools for Applied Multivariate Analysis.* New York: Academic Press.
10. Harris, R. J. 2001. *A Primer of Multivariate Statistics,* 3rd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.
11. Hosmer, D. W., and S. Lemeshow. 2000. *Applied Logistic Regression,* 2nd ed. New York: Wiley.
12. Huberty, C. J. 1984. Issues in the Use and Interpretation of Discriminant Analysis. *Psychological Bulletin* 95: 156–71.
13. Huberty, C. J., J. W. Wisenbaker, and J. C. Smith. 1987. Assessing Predictive Accuracy in Discriminant Analysis. *Multivariate Behavioral Research* 22 (July): 307–29.
14. Johnson, N., and D. Wichern. 2002. *Applied Multivariate Statistical Analysis,* 5th ed. Upper Saddle River, NJ: Prentice Hall.
15. Long, J. S. 1997. *Regression Models for Categorical and-Limited Dependent Variables: Analysis and Interpretation.* Thousand Oaks, CA:-Sage.
16. Morrison, D. G. 1969. On the Interpretation of Discriminant Analysis. *Journal of Marketing Research* 6(2): 156–63.
17. Pampel, F. C. 2000. *Logistic Regression: A Primer,* Sage University Papers Series on Quantitative Applications in the Social Sciences, # 07–096. Newbury Park, CA: Sage.
18. Perreault, W. D., D. N. Behrman, and G. M. Armstrong. 1979. Alternative Approaches for Interpretation of Multiple Discriminant Analysis in Marketing Research. *Journal of Business Research* 7: 151–73.