# Дискриминантный анализ

**Дискриминантный анализ для двух групп**

В маркетинговом исследовании «Выбор универмага» маркетологи использовали дискриминантный анализ для двух групп, чтобы выяснить, отличаются ли значения, которые присваивают восьми факторам выбора универмага респонденты, которым известны универмаги, от тех значений, которые выставляют респонденты, которым универмаги не известны.

Зависимой переменной были две группы респондентов — владеющие и не владеющие информацией об универмагах, а независимыми переменными — важность (значение) восьми факторов критерия выбора. Общая дискриминантная функция была значимой, что указывало на существенное различие между двумя группами. Результаты показали, что по сравнению с респондентами, ничего не знающими о данных универмагах, респонденты, хорошо осведомленные о них, придавали большее относительное значение качеству товаров, условиям возврата товаров, услугам продавцов, а также условиям кредитования и расчета с покупателями.

В указанном примере с универмагом фигурируют две группы респондентов (знакомые и не знакомые с универмагом), в то время как в примере с предрасположенностью к покупкам товаров со скидками проверяли три группы (лица, не являющиеся покупателями товаров со скидкой; редкие покупатели и частые). В данных исследованиях обнаружены существенные межгрупповые различия при использовании многих предикторов (независимых переменных). Исследование различий между группами — основа концепции дискриминантного анализа.

**Дискриминантный анализ (discriminant analysis)** используется для анализа данных втом случае, когда зависимая переменная категориальная, а предикторы (независимые переменные) интервальные.

**Дискриминантный анализ (discriminant analysis)**  
Метод для анализа данных маркетинговых исследований в том случае, когда зависимая переменная категориальная, а предикторы (независимые переменные) интервальные.

Например, зависимая переменная может быть выбором торговой марки персонального компьютера (торговые марки А, В или С), а независимыми переменными могут быть рейтинги свойств персональных компьютеров, измеренные по семибалльной шкале Лайкерта. Дискриминантный анализ преследует такие цели.

* Определение дискриминантных функций (discriminant functions) или линейных комбинаций независимых переменных, которые наилучшим образом различают (дискриминируют) категории (группы) зависимой переменной.
* Проверка существования между группами значимых различий с точки зрения независимых переменных.
* Определение предикторов, вносящих наибольший вклад в межгрупповые различия.
* Отнесение случаев к одной из групп (классификация) исходя из значений предикторов.
* Оценка точности классификации данных на группы.

Дискриминантная функция (discriminant functions)Выведенная с помощью дискриминантного анализа линейная комбинация независимых переменных, с помощью которой можно наилучшим образом различить (дискриминировать) категории зависимой переменной.

Метод дискриминантного анализа описывается количеством категорий, имеющихся у зависимой переменной. Если она имеет две категории, то метод называют дискриминант ным анализом для двух групп (two-group discriminant analsysis).

**Дискриминантный анализ для двух групп (two-group discriminant analysis)**Метод дискриминантного анализа, когда зависимая переменная имеет две категории. Если анализируют три или больше категорий, то метод называют множественным дискриминантным анализом (multiple descriminant analysis).

**Множественный дискриминантный анализ (multiple descriminant analysis)**  
Метод дискриминантного анализа, когда у зависимой переменной есть три или больше категорий.

Главное отличие между ними заключается в том, что при наличии двух групп можно вывести только одну дискриминантную функцию. Используя множественный дискриминантный анализ, можно вычислить несколько функций.

В маркетинговых исследованиях можно привести массу примеров применения дискриминантного анализа. Так, с помощью этого метода можно получить ответы на следующие вопросы.

* Чем, с точки зрения демографических характеристик, отличаются приверженцы данного магазина от тех, у кого эта приверженность отсутствует?
* Отличаются ли в потреблении замороженных продуктов покупатели, которые пьют безалкогольные напитки мало, умеренно и много?
* Какие психографические характеристики помогают провести различия между восприимчивыми и не восприимчивыми к цене покупателями бакалейных товаров?
* Различаются ли между собой различные сегменты рынка по своим предпочтениям к средствам массовой информации?
* Какие существуют различия между постоянными покупателями местных универсальных магазинов и постоянными покупателями общенациональных сетей универмагов с точки зрения стиля жизни?
* Какими отличительными характеристиками обладают потребители, реагирующие на прямую почтовую рекламу

**Модель дискриминантного анализа** (discriminant analysis model) имеет следующий вид:  
  
D = b0 + blXl + b2×2+b3Xi+...+bkXk,  
  
где D — дискриминантный показатель (дискриминант), b — дискриминантный коэффициент, или вес, X- предиктор, или независимая переменная.

**Модель дискриминантного анализа** (discriminant analysis mode). Статистическая модель, лежащая в основе дискриминантного анализа. Коэффициенты, или веса (D), определяют таким образом, чтобы группы максимально отличались значениями дискриминантной функции. Это происходит тогда, когда отношение межгрупповой суммы квадратов к внутригрупповой сумме квадратов для дискриминантных показателей максимально. Любая другая линейная комбинация предикторов приводит к меньшему значению этого отношения. С дискриминантным анализом связан ряд статистик.

Статистики, связанные с дискриминантным анализом

**Каноническая корреляция (canonical correlation)**. Измеряет степень связи между дискриминантными показателями и группами. Это мера связи между единственной дискриминирующей функцией и набором фиктивных переменных, которые определяют принадлежность к данной группе.

**Центроид (средняя точка) (centroid)**. Центроид — это средние значения для дискриминантных показателей конкретной группы. Центроидов столько, сколько групп, т.е. один центроид для каждой группы. Средние группы для всех функций — это групповые центроиды.

**Классификационная матрица (classification matrix)**. Иногда ее называют смешанной матрицей, или матрицей предсказания. Классификационная матрица содержит ряд правильно классифицированных и ошибочно классифицированных случаев. Верно классифицированные случаи лежат на диагонали матрицы, поскольку предсказанные и фактические группы одни и те же. Элементы, не лежащие на диагонали матрицы, представляют случаи, классифицированные ошибочно. Сумма элементов, лежащих на диагонали, разделенная на общее количество случаев, дает коэффициент результативности.

**Коэффициенты дискриминантной функции (discriminant function coefficients)**. Коэффициенты дискриминантной функции (ненормированные) — это коэффициенты переменных, когда они измерены в первоначальных единицах.

**Дискриминантные показатели (discriminant scores)**. Сумма произведений ненормированных коэффициентов дискриминантной функции на значения переменных, добавленная к постоянному члену.

**Собственное (характеристическое) значение (eigenvalue)**. Для каждойдискриминант-ной функции собственное значение — это отношение межгрупповой суммы квадратов к внутригрупповой сумме квадратов. Большие собственные значения указывают на функции более высокого порядка.

**F-статистика и ее значимость (F values and their significance)**. Значения f-статис-тики вычисляют с помощью однофакторного дисперсионного анализа, используя разбивку на группы независимой переменной. Каждый предиктор, в свою очередь, служит в ANOVA метрической зависимой переменной.

**Средние группы и групповые стандартные отклонения (group means and group standard deviations)**. Эти показатели вычисляют для каждого предиктора каждой группы.

**Объединенная межгрупповая корреляционная матрица (pooled within-group correlation matrix)**. Объединенную межгрупповую корреляционную матрицу вычисляют усреднением отдельных ковариационных матриц для всех групп.

**Нормированные коэффициенты дискриминантных функций (standardized discriminant function coefficients)**. Коэффициенты дискриминантных функций используют как множители для нормированных переменных, т.е. переменных с нулевым средним и дисперсией, равной 1.

**Структурные коэффициенты корреляции (structure correlations)**. Также известны как дискриминантные нагрузки, представляют собой линейные коэффициенты корреляции между предикторами и дискриминантной функцией.

**Общая корреляционная матрица (total correlation matrix)**. Если при вычислении корреляций наблюдения обрабатывают так, как будто они взяты из одной выборки, то в результате получают общую корреляционную матрицу.

**Коэффициент «лямбда» Уилкса (Wilks’s)**. Иногда называемый статистикой, коэффициент Уилкса для каждого предиктора — это отношение внутригрупповой суммы квадратов к общей сумме квадратов. Его значение варьируется от 0 до 1. Большое значение (около 1) указывает на то, что средние групп не должны различаться. Малые значения (около 0) указывают на то, что средние групп различаются.

В дискриминантном анализе существуют такие допущения: каждая группа является выборкой из многомерной нормально распределенной совокупности; все совокупности имеют одну и ту же ковариационную матрицу. Чтобы лучше понять роль допущений и описанных выше статистик, следует изучить методы выполнения дискриминантного анализа.

|  |
| --- |
| Выполнение дискриминантного анализа |
|  |

Выполнение дискриминантного анализа включает следующие стадии: формулирование проблемы, вычисление коэффициентов дискриминантной функции, определение значимости, интерпретация и проверка достоверности (рис. 1). Эти стадии обсуждаются и иллюстрируются для дискриминантного анализа двух групп.

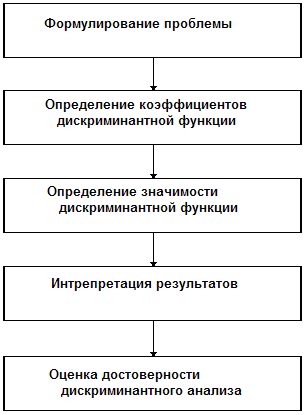


Рис. 1. Выполнение дискриминантного анализа.

**Формулирование проблемы**  
Первый шаг дискриминантного анализа — формулирование проблемы с помощью определения целей, зависимой переменой и независимых переменных. Зависимая переменная должна состоять из двух или больше взаимоисключающих и взаимно исчерпывающих категорий. Если зависимая переменная измерена с помощью интервальной или относительной шкалы, то ее следует в первую очередь перевести в статус категориальной. Например, отношение к торговой марке, измеренное по семибалльной шкале, можно категоризировать как неблагоприятное (1, 2, 3), нейтральное (4) и благоприятное (5, 6, 7). Можно поступить иначе. Для этого следует построить график распределения значений зависимой переменной и сформировать группы равного размера с помощью точек отсечения. Предикторы следует выбирать исходя из теоретической модели или уже проведенного исследования, или, в случае поискового исследования, из интуиции и опыта исследователя.

**Следующий шаг** — разделение выборки на две части. Одна из них — анализируемая выборка (analysis sample) — используется для вычисления дискриминантной функции.

**Анализируемая выборка (analysis sample)**Часть общей выборки, которую используют для вычисления дискриминантной функции.

**Другая часть** — проверочная выборка (validation sample) — предназначена для проверки дискриминантной функции.

**Проверочная выборка (validation sample)**Часть общей выборки, которую используют для проверки результатов расчета на основании анализируемой выборки.

Когда выборка достаточно велика, ее можно разбить на две равные части. Одна служит анализируемой выборкой, а другую используют для проверки. Затем роль этих половинок взаимно меняют и повторяют анализ. Это называется двойной перекрестной проверкой, и она аналогична методу, рассмотренному в регрессионном анализе.

Часто распределение количества случаев в анализируемой и проверочной выборках следует из распределения в общей выборке. Например, если общая выборка содержит 50% лояльно и 50% нелояльно настроенных покупателей, то анализируемая и проверочная выборки должны содержать каждая по 50% лояльных и 50% нелояльных покупателей. В другом случае, если выборка содержит 25% лояльных и 75% нелояльных покупателей, следует выбрать анализируемую и проверочную выборки таким образом, чтобы их распределения отражали аналогичную картину (25% против 75%).

И наконец, проверку достоверности дискриминантной функции предлагают выполнять неоднократно. Каждый раз выборку следует разбивать на две части — для анализа и проверки. Вычисляют дискриминантную функцию и выполняют анализ достоверности модели. Таким образом, оценка достоверности основана на ряде испытаний. Предлагаются также более точные методы.

Чтобы лучше проиллюстрировать дискриминантный анализ для двух групп, обратимся к примеру. Предположим, что мы хотим определить главные характеристики семей, которые отдыхали на курорте в последние два года. Данные получены на основании выборки, включающей 42 семьи. Из них 30 включены в анализируемую выборку, а оставшиеся 12 тали частью проверочной выборки.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| №п.п | Посещение курорта | Ежегодный доход семьи (тыс. долл.) | Отношение к путешествию | Значение, придаваемое семейному отдыху | Размер семьи | Возраст главы семьи | Сумма, потраченная семьей на отдых |
| 1 | 1 | 50,2 | 5 | 8 | 3 | 43 | С (2) |
| 2 | 1 | 70,3 | 6 | 7 | 4 | 61 | Б(3) |
| 3 | 1 | 62,9 | 7 | 5 | 6 | 52 | Б(3) |
| 4 | 1 | 48,5 | 7 | 5 | 5 | 36 | М(1) |
| 5 | 1 | 52,7 | 6 | 6 | 4 | 55 | Б(3) |
| 6 | 1 | 75,0 | 8 | 7 | 5 | 68 | Б(3) |
| 7 | 1 | 46,2 | 5 | 3 | 3 | 62 | С (2) |
| 8 | 1 | 57,0 | 2 | 4 | 6 | 51 | С (2) |
| 9 | 1 | 64,1 | 7 | 5 | 4 | 57 | Б(3) |
| 10 | 1 | 68,1 | 7 | 6 | 5 | 45 | Б(3) |
| 11 | 1 | 73,4 | 6 | 7 | 5 | 44 | Б(3) |
| 12 | 1 | 71,9 | 5 | 8 | 4 | 64 | Б(3) |
| 13 | 1 | 56,2 | 1 | 8 | 6 | 54 | С (2) |
| 14 | 1 | 49,3 | 4 | 2 | 3 | 56 | Б(3) |
| 15 | 1 | 62,0 | 5 | 6 | 2 | 58 | Б(3) |
| 16 | 2 | 32,1 | 5 | 4 | 3 | 58 | М(1) |
| 17 | 2 | 36,2 | 4 | 3 | 2 | 55 | М(1) |
| 18 | 2 | 43,2 | 2 | 5 | 2 | 57 | С (2) |
| 19 | 2 | 50,4 | 5 | 2 | 4 | 37 | С (2) |
| 20 | 2 | 44,1 | 6 | 6 | 3 | 42 | С (2) |
| 21 | 2 | 38,3 | 6 | 6 | 2 | 45 | М(1) |
| 22 | 2 | 55,0 | 1 | 2 | 2 | 57 | С (2) |
| 23 | 2 | 46,1 | 3 | 5 | 3 | 51 | М(1) |
| 24 | 2 | 35,0 | 6 | 4 | 5 | 64 | М(1) |
| 25 | 2 | 37,3 | 2 | 7 | 4 | 54 | М(1) |
| 26 | 2 | 41,8 | 5 | 1 | 3 | 56 | С (2) |
| 27 | 2 | 57,0 | 8 | 3 | 2 | 36 | С (2) |
| 28 | 2 | 33,4 | 6 | 8 | 2 | 50 | М(1) |
| 29 | 2 | 37,5 | 6 | 2 | 3 | 48 | М(1) |
| 30 | 2 | 41.3 | 3 | 3 | 2 | 42 | М(1) |

Семьям, которые отдыхали на курорте в последние два года, присвоен код 1; тем же, которые не посетили курорт за указанный период времени, присвоен код 2. Обе выборки (как анализируемая, так и проверочная) сбалансированы с точки зрения посещаемости курорта. Как видно, анализируемая выборка содержит 15 семей каждой категории, а проверочная — по 6 семей каждой категории. Кроме того, получены данные о ежегодном доходе каждой семьи (доход), отношении к путешествию (путешествие оценивали по девятибалльной шкале), значении, придаваемом семейному отдыху (отдых оценивали по девятибалльной шкале), размеру семьи (размер семьи) и возрасту главы семьи (возраст).

**Определение коэффициентов дискриминантной функции**  
После определения анализируемой выборки мы можем вычислить коэффициенты дискриминантной функции, используя два метода. Прямой метод (direct method) — вычисление дискриминантной функции при одновременном введении всех предикторов.

**Прямой метод** (direct method). Метод дискриминантного анализа, в котором дискриминантную функцию вычисляют при одновременном введении всех предикторов.

В этом случае учитывается каждая независимая переменная. При этом ее дискриминирующая сила не принимается во внимание. Этот метод больше подходит к ситуации, когда аналитик, исходя из результатов предыдущего исследования или теоретической модели, хочет, чтобы в основе различения лежали все предикторы. Альтернативным методом будет пошаговый метод. При пошаговом дискриминантом анализе (stepwise discriminant analysis) предикторы вводят последовательно, в зависимости от их способности различить (дискриминировать) группы.

**Пошаговый дискриминантный анализ** (stepwise discriminant analysis)Дискриминантный анализ, при котором предикторы вводятся последовательно, в зависимости от их способности различить группы.

Этот метод лучше применять в ситуации, когда исследователь хочет отобрать подмножество предикторов для включения их в дискриминатную функцию. Коэффициент X (f-статистика) и f-критерий для одномерной выборки с одной и 28 степенями свободы.

|  |  |  |  |
| --- | --- | --- | --- |
| Переменная | Коэффициент X Уилкса | Значение F | Значимость |
| Доход | 0,45310 | 33,80 | 0,0000 |
| Путешествие | 0,92479 | 2,277 | 0,1425 |
| Отдых | 0,82377 | 5,990 | 0,0209 |
| Размер семьи | 0,65672 | 14,64 | 0,0007 |
| Возраст | 0,95441 | 1,338 | 0,2572 |

**Структурная матрица**  
Объединенная корреляционная матрица между дискриминирующими переменными и каноническими дискриминантными функциями (переменные расположены в соответствии с размером корреляции внутри функции)

|  |  |
| --- | --- |
|  | Функция 1 |
| Доход | 0,82202 |
| Размер семьи | 0,54096 |
| Отдых | 0,34607 |
| Путешествие | 0,21337 |
| Возраст | 0,20922 |

Ненормированные коэффициенты канонической дискриминантной функции

|  |  |
| --- | --- |
| ***Группа*** | Функция 1 |
| Доход | 0,8476710Е-01 |
| Путешествие | 0.4964455Е-01 |
| Отдых | 0,1202813 |
| Размер семьи | 0,4273893 |
| Возраст | 0.2454380Е-01 |
| (Константа) | -7,975476 |

Некоторые результаты можно получить, изучив групповые средние и стандартные отклонения. Маркетологи обнаружили, что в деление совокупности на две группы самый большой вклад внесла переменная «Доход». Кроме того, оказалось, что переменная «Значение, придаваемое семейному отдыху», важнее для различения групп, чем переменная «Отношение к путешествию». По возрасту главы семьи две группы различаются мало, а стандартное отклонение этой переменной большое.

Объединенная внутригрупповая корреляционная матрица указывает на низкие коэффициенты корреляции между предикторами. Маловероятно, что возникнет проблема мультиколлинеарности. Значимость одномерных f-статистик (отношений внутригрупповых сумм квадратов к общей сумме квадратов) указывает, что когда предикторы рассматриваются по отдельности, то только доход, а также значение, придаваемое семейному отдыху, и размер семьи значимо различаются между семьями, которые посетили курорт, и между теми, кто не отдыхал на курорте.

Поскольку имеется две группы, оценивается только одна дискриминантная функция. Собственное значение, соответствующее этой функции, равно 1,7862. Каноническая корреляция, соответствующая этой функции, равна 0,8007. Квадрат корреляции, равный (0,8007)2 = 0,64, показывает, что 64% дисперсии зависимой переменной (посещение курорта) объясняется этой моделью. Следующая стадия дискриминантного анализа включает определение значимости дискриминантной функции.

**Определение значимости дискриминантной функции**  
Бессмысленно интерпретировать результаты анализа, если определенные дис-криминантные функции не будут статистически значимыми. Поэтому следует выполнить статистическую проверку нулевой гипотезы о равенстве средних всех дискриминантных функций во всех группах генеральной совокупности. В программе SPSS эта проверка базируется на коэффициенте X Уилкса. Если одновременно проверяют несколько функций, как в случае множественного дискриминантного анализа, то коэффициент X является суммой одномерных X для каждой функции. Уровень значимости оценивают исходя из преобразования F-статистики в статистику хи-квадрат (исходя из распределения X-квадрат, которому подчиняется F-статистика). При проверке значимости в примере с посещением курорта можно отметить, что X, равная 0,3589, преобразуется в хи-квадрат-статистику, равную 26,13 с пятью степенями свободы. Она значима при уровне, превышающем 0,05. В программе SAS вычисляют приближенную F-статистику, основанную на апроксимации к распределению отношения правдоподобия. В программе BMDP проверка нулевой гипотезы базируется на преобразовании Х- статистики Уилкса в F-статистику. В Minitab нельзя выполнить проверку значимости. Если нулевую гипотезу отклоняют, что указывает на значимую дискриминацию, то можно приступать к интерпретации результатов.

**Интерпретация результатов**  
Интерпретация дискриминантных весов, или коэффициентов, аналогична интерпретации результатов множественного регрессионного анализа. Значение коэффициента для конкретного предиктора зависит от других предикторов, включенных в дискриминантную функцию. Знаки коэффициентов условны, но они указывают, какие значения переменной приводят к большим и маленьким значениям функции, и связывают их с конкретными группами.

При наличии мультиколлинеарности между независимыми переменными не существует однозначной меры относительной важности предикторов для дискриминации между группами. Помня об этом предостережении, можно получить некоторое представление об относительной важности переменных, изучив абсолютные значения нормированных коэффициентов дискриминантной функции. Как правило, предикторы с относительно большими нормированными коэффициентами вносят больший вклад в дискриминирующую мощность функции по сравнению с предикторами, имеющими меньшие коэффициенты.

Некоторое представление об относительной важности предикторов можно также получить, изучив структурные коэффициенты корреляции, которые также называют **каноническими**, или **дискриминантными**,**нагрузками**. Эти линейные коэффициенты корреляции между каждым из предикторов и дискриминантной функцией представляют дисперсию, которую предиктор делит вместе с функцией. Как и нормированные коэффициенты, эти коэффициенты корреляции следует использовать осторожно.

Полезно исследовать нормированные коэффициенты дискриминантной функции в примере с отпуском на курорте. С данными низкими коэффициентами корреляции между предикторами можно использовать значения нормированных коэффициентов, чтобы предположить, что доход — наиболее важный предиктор при дискриминации между группами, а за ним следуют размер семьи и значение, придаваемое семейному отдыху. Аналогичное наблюдение получено из проверки структурных корреляций. Эти коэффициенты линейной корреляции между предикторами и дискриминантной функцией перечислены в порядке их убывания.

Также даны и ненормированные коэффициенты дискриминантной функции. Для классификации данных их можно применить к необработанным значениям переменных в проверочной выборке. Кроме того, показаны групповые центроиды, дающие значения дискриминантной функции, оцененные по групповым средним. Центроид группы 1 (семьи, отдыхающие на курорте) имеет положительное значение, а центроид группы 2 — равное ему, но отрицательное. Знаки коэффициентов соответствующих предикторов положительны. Это означает, что чем выше доход семьи, ее размер, значение, придаваемое семейному отдыху, а также отношение к путешествию и возраст, тем выше вероятность семейной поездки на курорт. Разумно создать профиль двух групп с точки зрения трех предикторов, которые кажутся наиболее важными: доход, размер семьи и значение, придаваемое семейному отдыху.

При интерпретации результатов дискриминантного анализа также может помочь разработка**характеристической структуры** (characteristic profile) для каждой группы с помощью описания каждой группы через групповые средние для предикторов.

**Характеристическая структура** (characteristic profile). Средство интерпретации результатов дискриминантного анализа описанием каждой группы через групповые средние для предикторов.

Если важные предикторы установлены, то сравнение групповых средних по этим переменным поможет понять межгрупповые различия. Однако прежде чем интерпретировать какие-либо факты, необходимо убедиться в достоверности результатов.

**Оценка достоверности дискриминантного анализа**  
Как уже говорилось, данные разбивают случайным образом на две подвыборки. Анализируемую часть выборки используют для вычисления дискриминантной функции, а проверочную — для построения классификационной матрицы. Дискриминантные веса, определенные анализируемой выборкой, умножают на значения независимых переменных в проверочной выборке, чтобы получить дискриминантные показатели для случаев в этой выборке. Затем случаи распределяют по группам исходя из дискриминантных показателей и соответствующего правила принятия решения. Например, при дискриминантном анализе двух групп случай может быть отнесен к группе с самым близким по значению центроидом. Затем, сложив элементы, лежащие на диагонали матрицы, и разделив полученную сумму на общее количество случаев, можно определить **коэффициент результативности** (hit ratio), или процент верно классифицированных случаев.

**Коэффициент результативности** (hit ratio). Процент случаев, верно классифицированных с помощью дискриминантного анализа. Полезно сравнить процент случаев, верно классифицированных с помощью дискриминантного анализа, с процентом случаев, который можно получить случайным образом. Для равных по размеру групп процент случайной классификации равен частному от деления единицы на количество групп. Превысит ли и на сколько количество верно классифицированных случаев их случайное количество? Здесь нет общепринятого подхода, хотя некоторые считают, что точность классификации, достигнутая с помощью дискриминантного анализа, должна быть, по крайней мере, на 25% выше, чем точность, которую можно достичь случайным образом.

Многие программы для выполнения дискриминантного анализа также определяют классификационную матрицу исходя из анализируемой выборки. Поскольку программы учитывают даже случайные вариации в данных, полученные результаты всегда точнее, чем классификация данных на основе проверочной выборки.

Коэффициент результативности, или процент верно классифицированных случаев, равен (12 + 15)/30 = = 0,90, или 90%. Могут возникнуть сомнения, что этот коэффициент результативности искусственно завышен, поскольку данные, использованные для вычисления, использовались и для проверки. Выполнение классификационного анализа по независимому набору данных приводит к классификационной матрице с немного меньшим коэффициентом результативности (4 + 6)/12 = 0,833, или 83,3%. Задав случайным образом две группы равного размера, можно ожидать, что коэффициент результативности равен 1/2 = 0,50, или 50%. Однако превышение точности классификации над случайной классификацией составляет свыше 25%, и поэтому достоверность дискриминантного анализа оценивают как удовлетворительную.

|  |
| --- |
| Пошаговый дискриминантный анализ |
|  |

Пошаговый дискриминантный анализ аналогичен пошаговому множественному регрессионному анализу в том отношении, что предикторы вводят последовательно, исходя из их способности различать (дискриминировать) группы. Значение F-статистики рассчитывают для каждого предиктора, выполняя одномерный дисперсионный анализ, в котором группы рассматривают как категориальную переменную, а предиктор — как критериальную переменную. Предиктор с самым высоким значением F-статистики первым отбирают для включения в дискриминантную функцию, если он удовлетворяет определенной значимости и допустимому критерию. Второй предиктор вводят исходя из самого высокого скорректированного или частного значения F, с учетом уже выбранного предиктора.

Для того чтобы каждый выбранный предиктор оставить в уравнении, его проверяют исходя из его связи с другими предикторами. Процесс введения и исключения продолжают до тех пор, пока все предикторы не будут удовлетворять критерию значимости — условию, необходимому для их введения в дискриминантную функцию. На каждой стадии рассчитывают несколько статистик. Кроме того, в заключение подводят итог введенным или исключенным предикторам. Пошаговый метод приводит к тому же стандартному выводу, который вытекает из прямого метода.

Выбор пошагового метода основан на оптимизации принятого критерия. Метод Махаланобиса (Mahalanobis procedure) основан на максимизации обобщенной меры расстояния между двумя самыми близкими группами. Этот метод позволяет маркетологам-исследователям извлечь максимальную пользу из имеющейся информации.

**Метод Махаланобиса (Mahalanobis procedure)**  
Пошаговый метод, используемый в дискриминантном анализе для максимизации обобщенной меры расстояния между двумя самыми близкими группами.

Первой выбранной переменной был доход, за ним следовали размер семьи и отдых. Порядок введения переменных указывает на их значимость в дискриминации двух групп. Впоследствии это подтвердила проверка нормированных коэффициентов дискриминантной функции и структура коэффициентов корреляции. Обратите внимание, что результаты пошагового анализа согласуются с выводами, ранее полученными прямым методом.

Врезки «Практика маркетинговых исследований» — примеры применения дискриминантного анализа в международных маркетинговых исследованиях и при исследование этических проблем в маркетинге.