СТАТИСТИЧЕСКИЕ МЕТОДЫ И АНАЛИЗ ДАННЫХ

Г.П. Бессокирная
(*Москва*)

ДИСКРИМИНАНТНЫЙ АНАЛИЗ ДЛЯ ОТБОРА ИНФОРМАТИВНЫХ ПЕРЕМЕННЫХ

В статье описываются возможности дискриминантного анализа, который позволяет выявлять различия между априорно заданными группами объектов по нескольким переменным одновременно, а также распознавать новые объекты по принципу максимального сходства. На конкретном примере обсуждаются результаты.

Ключевые слова: дискриминантная функция, группирующая переменная, дискриминантные переменные, пошаговый дискриминантный анализ, статистика Уилкса, расстояние Махалонобиса.

Дискриминантный анализ (ДА) – комплекс методов многомерного статистического анализа, используемых в определенной последовательности. Суть заключается в следующем. Имеются качественно разнородные априорно заданные группы объектов (так называемая обучающая выборка). ДА позволяет «построить» одну или несколько дискриминантных функций, которые наилучшим образом характеризуют различия между этими группами объектов и могут использоваться.

Теоретические основы ДА описаны в справочной литературе и в учебниках по прикладной статистике [например, 1; 2]. В качестве отправной точки для знакомства с этим разделом многомерного

Галина Петровна Бессокирная – кандидат экономических наук, старший научный сотрудник Института социологии РАН. E-mail: gala@isras.ru.

статистического анализа будут весьма полезны работа У. Клекка [3] и пособия по SPSS, опубликованные на русском языке [4; 5; 6]. В этих публикациях подробно изложены и цели, для достижения которых применяется дискриминантный анализ в социальных науках. В отечественной социологии в настоящее время ДА используется крайне редко, поэтому прежде всего остановимся кратко на описании методов ДА и особенностях интерпретации результатов его применения.

Кратко о дискриминантном анализе

Дискриминантный анализ – это общий термин для обозначения логики совместного использования нескольких тесно связанных статистических процедур. Как уже отмечалось, они дают возможность изучать межгрупповые различия по нескольким переменным одновременно в априорно заданных группах и распознавать (классифицировать) новые объекты.

Переменную, которую исследователь применяет для разделения объектов на группы, называют группирующей переменной. ДА целесообразно использовать, когда уровень измерения группирующей переменной номинальный или порядковый. Если речь идет о более высоком уровне измерения, то имеет смысл обратиться к регрессионному анализу.

Переменные, которыми оперируют для поиска различий между группами, называют дискриминантными переменными. Теоретически они должны иметь более высокий уровень измерения, чем группирующая переменная, но на практике используют все типы переменных. Например, номинальные переменные преобразуют в дихотомические переменные. К последним переменным допустимо применение любых количественных методов [7, с. 306–310].

В ДА не делается предположений о зависимости или независимости группирующей переменной и дискриминантных переменных. Если в конкретной ситуации группирующую переменную

можно считать зависимой от дискриминантных переменных, то задача ДА аналогична задаче множественной регрессии. Основное отличие в том, что в ДА, как подчеркивалось выше, группирующая переменная не является количественной. В случае, когда предполагается наличие зависимости значений дискриминантных переменных от принадлежности к группе, ДА является аналогом многомерного дисперсионного анализа [3, с. 83].

На первом этапе проведения ДА изучаются межгрупповые различия и ищутся ответы на следующие вопросы: возможно ли, используя заданный набор дискриминантных переменных, отличить одну группу от другой; насколько хорошо эти переменные позволяют провести различение; какие из них наиболее информативны.

Одним из способов отбора информативных дискриминантных переменных является пошаговый дискриминантный анализ. Логика пошагового ДА такова: вначале определяется та переменная, для которой средние значения в априорно заданных группах «наиболее различны». На каждом следующем шаге рассматриваются условные распределения оставшихся переменных и определяется та, для которой средние значения в группах «наиболее различны». Процесс завершается, когда ни одна из оставшихся переменных не вносит значимого вклада в различение групп [8, с. 344–345].

Существуют разные критерии отбора информативных дискриминантных переменных. Каждый критерий придает особый содержательный смысл процессу различения. Наиболее популярная – это статистика Уилкса [5, с. 259], которая учитывает как различия между группами, так и однородность каждой из групп [3, с. 124]. Во многих исследованиях для отбора информативных переменных используют расстояние Махалонобиса [5, с. 259–260]. Эта статистика выделяет переменную, которая порождает наибольшее различение пары групп, являющихся ближайшими на данном шаге [3, с. 125]. В ряде случаев результат ДА не зависит от выбора критерия отбора дикриминантных переменных.

Процедура пошагового ДА предполагает также проверку (в начале каждого шага) всех дискриминантных переменных на соответствие двум условиям: необходимой точности вычисления (толерантности) и превышению заданного уровня различения (с этой целью используют статистики *F*-ввода и *F*-исключения). Статистика *F*-ввода оценивает улучшение различения благодаря использованию данной переменной по сравнению с различением, достигнутым с помощью уже отобранных переменных. Статистика *F*-исключения определяет значимость ухудшения различения после удаления переменной из списка уже отобранных переменных. На заключительном шаге статистика *F*-исключения может быть использована для оценки дискриминантных возможностей отобранных переменных. Переменная с наибольшим значение *F*-исключения дает наибольший вклад в различение, достигнутое посредством других переменных. Переменная, имеющая вторую по величине статистику *F*-исключения, является второй по значимости и т.д.

На втором этапе проведения ДА отобранное подмножество наиболее информативных переменных используется для вычисления дискриминантных функций. Дискриминантная функция является линейной комбинацией дискриминантных переменных и выглядит как правая часть уравнения множественной регрессии. Исследователь получает одну или несколько дискриминантных функций. Для двух априорно заданных групп вычисляется одна дискриминантная функция. Если групп более двух, то вычисляются несколько дискриминантных функций, которые не коррелированы между собой (число таких функций равно числу групп минус 1). Они называются каноническими дискриминантными функциями (каноническими переменными)[1].

[1] В работе У. Клекка показано, что в этом случае ДА является аналогом канонического корреляционного анализа [3, с. 107]. Отсюда и название канонические дискриминантные функции (канонические переменные).

На наш взгляд, для социолога важны не столько принципы вычисления дискриминантных функций, сколько особенности содержательной интерпретации результатов применения ДА.

1. Прежде всего в процессе интерпретации исследователь решает: все ли вычисленные канонические дискриминантные функции полезны для описания межгрупповых различий. С этой целью используется собственное значение и относительное процентное содержание вычисленных функций (% объясненной дисперсии), коэффициент канонической корреляции[1], а также тест равенства средних значений канонических дискриминантных функций в группах. Опыт применения ДА показал, что о реальной полезности канонических дискриминантных функций для различения объектов целесообразнее всего судить по величине коэффициента канонической корреляции [3, с. 108]. Если коэффициент – величина небольшая, то каноническую дискриминантную функцию (каноническую переменную) обычно для интерпретации не используют.

2. На относительный вклад отдельных дискриминантных переменных в значение каждой дискриминантной функции указывают стандартизированные коэффициенты. Соответственно, чем больше стандартизированный коэффициент, тем больше вклад переменной.

3. О тесноте связи между отдельными дискриминантными переменными и дискриминантными функциями можно судить по величине структурных коэффициентов. Когда абсолютная величина такого коэффициента велика, вся информация о дискриминантной функции заключается в этой переменной. Исходя из тех структурных коэффициентов, которые максимальны по абсолютной величине, можно дать «имя» дискриминантной функции [3, с. 108].

---

[1] Коэффициент канонической корреляции – мера связи между априорно заданной принадлежностью к группе и вычисленной канонической дискриминантной функцией. Он идентичен смешанному моменту корреляции Пирсона между двумя линейными комбинациями в паре [3, с. 106].

Самым лучшим показателем информативности отобранных дискриминантных переменных и полезности применения дискриминантной функции для интерпретации межгрупповых различий является, конечно, процент правильно распознанных объектов с использованием вычисленных дискриминантных функций (классификация с учителем). Число правильно распознанных новых объектов (как в целом, так и по отдельным группам) свидетельствует о соответствии дискриминантной модели эмпирическим данным.

Таким образом, ДА в социологических исследованиях можно применять как для интерпретации межгрупповых различий, так и для классификации новых объектов. Часто социолог оказывается в ситуации, когда в его распоряжении имеется несколько дискриминантных переменных, и он пытается найти и исключить из дальнейшего анализа наименее информативные переменные. ДА позволяет выделить некоторое подмножество дискриминантных переменных, которое столь же эффективно для различения групп, как и все множество дискриминантных переменных.

Дискриминантный анализ факторов социальной адаптации рабочих

В исследовании социальной адаптации рабочих нашей задачей являлось выяснение относительной значимости для формирования стратегий выживания: ресурсного потенциала работников и их домохозяйств; их места работы. Для ее решения посредством применения ДА был сделан вывод, что наиболее значимы для объяснения и прогнозирования стратегий выживания рабочих такие показатели, как их возраст, доля работающих в семье и тип предприятия [9, с. 114]. Ниже рассмотрим логику решения этой задачи.

Группы стратегий выживания были выделены на основе анализа данных российского мониторинга экономического положения и здоровья населения о распространенности и эффективности различных способов приспособления занятого городского насе-

ления к новым условиям жизни. Выяснилось, что относительно эффективными (по оценкам самих респондентов) в 1990-е гг. были активные способы приспособления, в первую очередь, интенсификация работ на своем приусадебном участке и дополнительная работа. Всего было выделено четыре стратегии выживания:

новаторская стратегия – ориентация или участие во вторичной занятости плюс отсутствие садово-огородного участка или желания его иметь;

смешанная стратегия – вторичная занятость или ориентация на такую занятость плюс наличие садово-огородного участка или желания его иметь;

традиционная стратегия – садово-огородный участок и отсутствие вторичной занятости;

пассивная стратегия – отсутствие вторичной занятости и садово-огородного участка [10].

В обучающей выборке общим объемом 597 человек рабочие были распределены на четыре группы следующим образом: смешанная стратегия –35%, новаторская стратегия – 26%, традиционная стратегия – 22%, пассивная стратегия – 17%.

Итак, в качестве группирующей переменной была взята «стратегия выживания». В исходную совокупность дискриминантных переменных были включены: тип предприятия (частное, акционерное, государственное) и переменные, характеризующие ресурсный потенциал как самих рабочих (пол, возраст, стаж работы на предприятии), так и их домохозяйств (доля работающих в семье, доход на 1 члена семьи). Отбор именно этих переменных был осуществлен, исходя из теоретической модели социальной адаптации, а также результатов анализа коэффициентов парной связи между группирующей переменной и потенциальными дискриминантными переменными. Наш опыт проведения ДА показал, что успеху интерпретации межгрупповых различий по нескольким переменным одновременно может способствовать применение для такого отбора не только симметричных (вычислялся ко-

эффициент Крамера), но и направленных коэффициентов связи (вычислялись λ Л. Гуттмана и корреляционное отношение).

Обработка данных осуществлялась с помощью пакета SPSS. Было апробировано три варианта пошагового ДА с использованием в качестве критерия отбора: статистики Уилкса (ДА-1), расстояния Махалонобиса (ДА-2), расстояния Махалонобиса для случая предварительно сгруппированных количественных дискриминантных переменных (ДА-3).

Посредством ДА-1 были определены только две переменные: возраст и доля работающих в семье. Тип предприятия не был отобран в качестве информативной переменной.

Упомянутый выше вывод о том, что для формирования стратегий выживания наиболее значимыми являются возраст, доля работающих в семье и место работы, был сделан по рузультатам ДА-2.

Дальнейшие эксперименты показали, что для успеха интерпретации межгрупповых различий по нескольким переменным одновременно весьма полезно сгруппировать данные. Например, данные о возрасте были интервалированы следующим образом: до 24 лет, 25-30 лет, 31-50 лет, старше 50 лет (с учетом особенностей социализации в различных условиях); данные о стаже работы на предприятии: поступил на работу до 1992 г. или в 1992 г. и позже (выбор места работы в советское время или на формирующемся рынке труда) и т.п.

Когда все количественные дискриминантные переменные были классифицированы, то в результате применения ДА-3 были отобраны в качестве наиболее информативных те же три переменные, что и при использовании ДА-2, но их значимость несколько изменилась. Наиболее значимыми стали: возраст, тип предприятия, доля работающих в семье. Результаты ДА-3 показывают, что две первые (по значимости) дискриминантные переменные позволяют наилучшим образом различать наиболее массовые априорно заданные группы (новаторскую и смешанную стратегии выживания), а третья переменная – две другие группы (традиционную и пассивную).

Судя по результатам всех трех вариантов пошагового ДА, определенный интерес для интерпретации межгрупповых различий представляют только первые канонические переменные, хотя фактически по всем каноническим переменным наблюдаются статистически значимые различия между группами (табл. 1).

Таблица 1

РЕЗУЛЬТАТЫ ДИСКРИМИНАНТНОГО АНАЛИЗА

	ДА-1		ДА-2			ДА-3		
	Номер канонической переменной							
	1	2	1	2	3	1	2	3
Собственное значение	0,186	0,027	0,191	0,044	0,009	0,166	0,028	0,014
% объясненной дисперсии	87,2	12,8	78,2	18,1	3,7	79,7	13,3	6,9
Коэффициент канонической корреляции	0,396	0,163	0,401	0,206	0,094	0,377	0,164	0,119
Уровень значимости	0,000	0,008	0,000	0,001	0,073	0,000	0,008	0,029

Первая каноническая переменная в ДА-1 объясняет 87,2% изменчивости исходных данных, в ДА-2 – 78,2%, а в ДА-3 – 79,7%. Процент объясненной дисперсии вторыми и третьими каноническими переменными невелик и, главное, коэффициенты канонической корреляции первых канонических переменных близки к 0,4, вторых – около 0,2, а третьих – около 0,1.

Сравнение результатов ДА-2 и ДА-3 (табл.1) свидетельствует о том, что при группировке количественных переменных (ДА-3) третья каноническая переменная имеет не только несколько больший процент объясненной дисперсии, но и несколько большую каноническую корреляцию. Исходя из этого и учитывая итоги пошагового ДА, результаты ДА-3 были выбраны для содержательной интерпретации канонических переменных (табл. 2, 3).

По величине стандартизированных коэффициентов (табл. 2) можно сделать вывод о том, что наибольший относительный вклад в значение первой канонической переменной (при фиксировании остальных) вносят тип предприятия и возраст рабочих.

Анализ структурных коэффициентов, которые отражают вклад каждой дискриминантной переменной в первую каноническую переменную (без учета ее коррелированности с другими), свидетельствует о том, что первая каноническая переменная тесно связана не только с типом предприятия и возрастом рабочих, но и со стажем работы на предприятии и полом (табл. 3). Эта каноническая переменная может быть названа «личный ресурс рабочего и его место работы».

Таблица 2

СТАНДАРТИЗИРОВАННЫЕ ДИСКРИМИНАНТНЫЕ КОЭФФИЦИЕНТЫ (РЕЗУЛЬТАТЫ ДА-3)

Наиболее информативные переменные	Канонические переменные		
	1	2	3
Тип предприятия	0,508	–0,964	–0,448
Возраст (4 группы)	0,507	0,955	–0,243
Доля работающих в семье (4 группы)	0,344	0,172	0,998

Таблица 3

ВНУТРИГРУППОВЫЕ СТРУКТУРНЫЕ КОЭФФИЦИЕНТЫ (РЕЗУЛЬТАТЫ ДА-3)

Дискриминантные переменные	Канонические переменные		
	1	2	3
Тип предприятия	0,842*	–0,499	–0,204
Возраст (4 группы)	0,749*	0,559	–0,355
Стаж работы на предприятии (2 группы)	0,673*	–0,146	–0,288
Пол	–0,635*	0,197	0,102
Доля работающих в семье (4 группы)	0,560	–0,088	0,824*
Доход на 1 члена семьи (3 группы)	–0,067	–0,191	0,273*

*Максимальные структурные коэффициенты для дискриминантных переменных.

В заключение следует отметить, что дискриминантный анализ позволил уточнить выводы о различиях в априорно заданных группах рабочих. В частности, было установлено, что наиболее

информативными переменными являются возраст и место работы. Именно эти переменные были использованы в процессе дальнейшего анализа социальной адаптации рабочих.

ЛИТЕРАТУРА

1. *Мешалкин Л.Д.* Теоретические результаты классификации при наличии обучающих выборок (дискриминантный анализ) // Прикладная статистика: Классификация и снижение размерности. М.: Финансы и статистика, 1989.

2. *Дубров А.М., Мхитарян В.С., Трошин Л.И.* Многомерные статистические методы. М.: Финансы и статистика, 2000.

3. *Клекка У.Р.* Дискриминантный анализ // Факторный, дискриминантный и кластерный анализ. М.: Финансы и статистика, 1989.

4. Дискриминантный анализ // SPSS. Base 7.5 для Windows: Руководство пользователя. М.: Стасис, 1997.

5. Дискриминантный анализ // SPSS. Base 7.5 для Windows: Руководство по применению. М.: Стасис, 1997.

6. *Бююль А., Цефель П.* SPSS: искусство обработки информации: Анализ статистических данных и восстановление скрытых закономерностей / Пер. с нем.; Под ред. В.Е. Момота. СПб.: ООО «ДиаСофтЮП», 2002.

7. *Толстова Ю.Н.* Анализ социологических данных. М.: Научный мир, 2000.

8. *Афифи А., Эйзен С.* Статистический анализ: Подход с использованием ЭВМ. М.: Мир, 1982.

9. *Бессокирная Г.П., Темницкий А.Л.* Социальная адаптация рабочих в трансформирующемся обществе: основные положения программы и некоторые результаты исследования // Мир России. 2000. № 4.

10. *Бессокирная Г.П.* Стратегии выживания горожан и их социальные последствия // Образ жизни горожан в объективных и субъективных показателях / Отв. ред. Т.М. Караханова. М.: Изд-во Института социологии РАН, 2002.