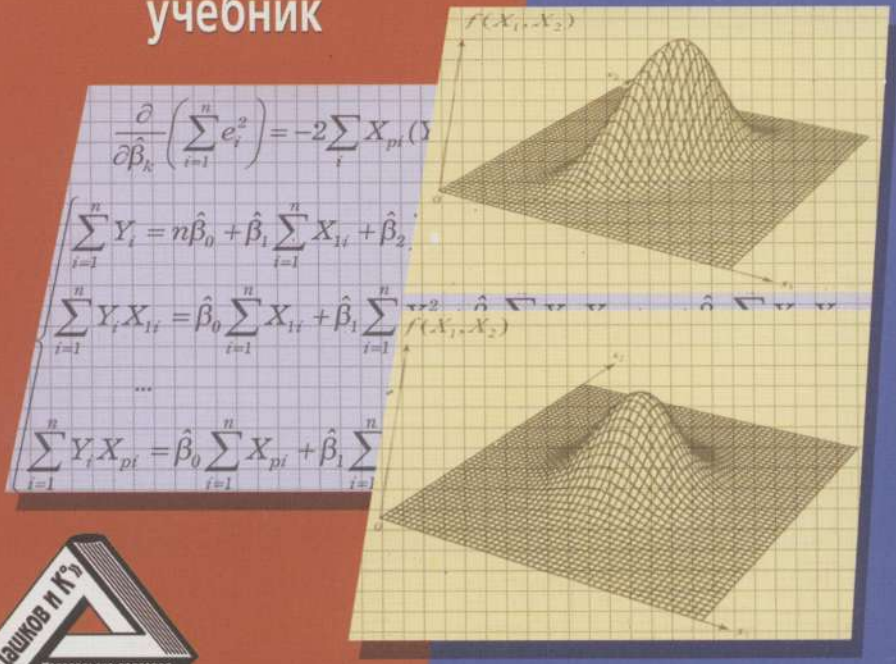
И. Ниворожкина, С. В. Арженовский

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ В ЭКОНОМИКЕ

учебник



«Дашков и К°»
Издательско-торговая корпорация

УДК 311
ББК 60.6
Н60

Рецензенты:
Кафедра прикладной математики Южно-Российского государственного технического университета (НПИ);
З. А. Морозова — кандидат экономических наук, доцент.

Н60 **Ниворожкина Л. И.**
Многомерные статистические методы в экономике: Учебник / Л. И. Ниворожкина, С. В. Арженовский. — М.: Издательско-торговая корпорация «Дашков и К°»; Ростов н/Д: Наука-Спектр, 2009. — 224 с.

ISBN 978-5-394-00469-8

Содержание учебника соответствует Государственному образовательному стандарту. Рассмотрены методы корреляционного, регрессионного, компонентного, факторного анализа для многомерной выборочной совокупности. Изложены методы классификации: дискриминантный и кластерный анализ. Особое внимание уделено ситуациям, при которых применение методов многомерного статистического анализа является некорректным.

Приведены примеры решения задач из экономической области. Учебник содержит необходимые сведения для выполнения прикладного многомерного анализа в программном комплексе Statistica.

Для студентов высших учебных заведений, обучающихся по специальностям «Статистика», «Математические методы в экономике» и другим экономическим специальностям, а также для аспирантов и научных работников, применяющих в исследованиях методы многомерного статистического моделирования.

УДК 311
ББК 60.6

ISBN 978-5-394-00469-8

СОДЕРЖАНИЕ

Предисловие

Смотри в корень!
К. Прутков. Мысли и афоризмы

В последнее время специалисты, обладающие знаниями и навыками проведения прикладного экономического анализа с использованием доступных математических и программных средств, пользуются повышенным спросом на рынке труда. Одной из основных дисциплин в их подготовке является курс «Многомерные статистические методы», дающий представление о многомерных случайных величинах и методах их анализа. Многомерные статистические методы являются областью знаний, которая охватывает вопросы применения статистических методов для поиска и объяснения закономерностей экономических процессов и явлений, скрытых от непосредственного наблюдения, позволяет структурировать и представлять в «сжатом» виде огромные информационные массивы, анализ которых методами традиционной статистики малоэффективен.

Расчеты в многомерном статистическом анализе проводят с помощью компьютеров. Существует широкий спектр пакетов прикладных программ, позволяющих автоматизировать процессы такого анализа. К наиболее распространенным относятся пакеты Statistica, SPSS, Stata, SAS и др. Имеются простейшие опции для проведения многомерного анализа в Excel. Поскольку в настоящее время вычислительные сложности анализа преодолены, а программные пакеты имеют, как правило, дружественный к пользователю интерфейс, то наиболее важным для осуществления прикладного многомерного анализа статистических данных представляется понимание сущности математико-статистических подходов, лежащих в их основе, выбор адекватной данным модели и после расчета ее параметров, интерпретация полученных результатов, дающая путь к объяснению экономической ситуации или принятию решений.

В книге даны основные понятия, модели и методы многомерного анализа, рассматриваются примеры из практики. В конце каждой главы приведены контрольные вопросы и задания, а также методические указания по выполнению расчетов в пакете прикладных программ Statistica. Содержание учебника полностью соответствует требованиям государственного стандарта высшего профессионального образования.

Для работы с предлагаемым изданием необходимы базовые знания некоторых разделов следующих учебных дисциплин: высшая математика (линейная алгебра, аналитическая геометрия), теория вероятностей, математическая статистика.

Эффективным является использование данной книги в сочетании с самостоятельным разбором примеров с использованием статистического программного обеспечения Statistica (наборы данных доступны в сети Интернет по адресу: ashad17.narod.ru).

Авторы благодарят рецензентов за советы при подготовке книги.

ГЛАВА 1
Содержание и основные этапы многомерного статистического анализа

Многие вещи нам непонятны не потому, что наши понятия слабы; но потому, что сии вещи не входят в круг наших понятий.

К. Прутков. Мысли и афоризмы

1.1. Задачи и методы многомерного статистического анализа

Многомерный статистический анализ (*МСА*) — раздел математической статистики, посвященный методам сбора, систематизации, обработки и интерпретации сложных совокупностей данных, нацеленный на выявление неявных (латентных) закономерностей в структуре и тенденциях развития исследуемых многомерных процессов.

Например, изучая модели экономического поведения человека, мы можем судить о нем по заработной плате и образованию, но наши выводы будут полнее и точнее, если мы включим в анализ такие признаки, как социальное положение, состав семьи, уровень доходов семьи, состояние здоровья и др. Совместное изучение значений этих признаков позволит

адекватно моделировать поведенческие реакции личности, коллектива.

Наиболее распространенными формами представления исходных статистических данных в МСА являются:

а) матрица объект-свойство

$$\underset{n\times p}{\mathbf{X}} = \begin{bmatrix} X_{11}(t) & X_{12}(t) & \dots & X_{1p}(t) \\ X_{21}(t) & X_{22}(t) & \dots & X_{2p}(t) \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1}(t) & X_{n2}(t) & \dots & X_{np}(t) \end{bmatrix}, \; t=t_1, \dots, t_N, \qquad (1.1)$$

где $X_{ij}(t)$ — значение j-го анализируемого признака, характеризующего состояние i-го объекта в момент времени t.

Например, пространственно-временная выборка, пространственная выборка при фиксированном t, временные ряды (n=1);

б) матрица парных сравнений.

Состоит из характеристик попарных сравнений объектов по некоторому свойству:

$$\underset{n\times n}{\Gamma} = \begin{bmatrix} \gamma_{11} & \dots & \gamma_{1n} \\ \vdots & \ddots & \vdots \\ \gamma_{n1} & \dots & \gamma_{nn} \end{bmatrix}. \qquad (1.2)$$

Содержание многомерного статистического анализа состоит в решении следующих основных проблем.

1. *Статистическое исследование зависимостей.* Выявление и описание множественных статистических связей, существующих между множеством признаков $X=(X_1, X_2, \dots, X_p)$. Используемые методы: корреляционный, регрессионный анализ, анализ временных рядов и др.

2. *Классификация объектов и признаков.* Необходимо всю анализируемую совокупность объектов $O_1, \dots, O_n$, представленную в виде матриц (1.1) или (1.2), разбить на сравнительно небольшое число (известное заранее или нет) однородных в определенном смысле групп или классов. Исходными данными при классификации объектов являются строки матрицы (1.1) или (1.2), при классификации признаков — столбцы матрицы (1.1) или (1.2).

Методы: дискриминантный анализ, кластерный анализ и др.

3. *Снижение размерности анализируемого признакового пространства.*

Переход от исходного набора из p признаков к вспомогательному набору меньшего числа P признаков. Это необходимо при решении задач: отбора наиболее информативных показателей, сжатия больших массивов информации, визуализации многомерных данных.

Методы: факторный анализ, метод главных компонент, многомерное шкалирование.

Основные этапы многомерного статистического анализа [1].

I. Предварительный анализ исследуемой реальной системы. Результат: определение цели и задач исследования, выбор объектов и признаков, формы для сбора информации, оценка необходимого времени и трудозатрат на проведение исследования.

II. Составление детального плана сбора исходной статистической информации.

III. Сбор и контроль исходных статистических данных и их преобразование в электронную форму.

IV. Первичная статистическая обработка данных.

Задачи:

- отображение вербальных признаков в номинальной или порядковой шкале,
- статистическое описание исходных совокупностей,
- анализ выбросов,
- восстановление пропущенных наблюдений,
- проверка однородности выборки,
- проверка статистической независимости последовательности наблюдений, составляющих выборку,
- экспериментальный анализ закона распределения исследуемой генеральной совокупности и др.

V. Уточнение методов анализа, используемых для моделирования исследуемой проблемы. Составление детального плана вычислительного анализа информации.

VI. Вычислительная реализация основной части статистической обработки данных.

VII. Подведение итогов исследования, интерпретация результатов, выводы.

На практике все перечисленные этапы не обязательно присутствуют и четко не разграничены. Некоторые из них могут объединяться или исключаться. Знание всех этапов позволяет рационально планировать реализацию методов МСА и учитывать предстоящие объемы работы.

Таблица 1.1

Взаимосвязь между методами многомерного исследования зависимостей

Метод	Вид зависимости
Канонические корреляции	$\underset{\text{(количественные, неколичественные)}}{Y_1 + Y_2 + \ldots + Y_m} = \underset{\text{(количественные, неколичественные)}}{X_1 + X_2 + \ldots + X_m}$
Многомерный дисперсионный анализ	$\underset{\text{(количественные)}}{Y_1 + Y_2 + \ldots + Y_m} = \underset{\text{(неколичественные)}}{X_1 + X_2 + \ldots + X_m}$
Дисперсионный анализ	$\underset{\text{(количественная)}}{Y_1} = \underset{\text{(неколичественные)}}{X_1 + X_2 + \ldots + X_m}$
Дискриминантный анализ	$\underset{\text{(неколичественная)}}{Y_1} = \underset{\text{(количественные)}}{X_1 + X_2 + \ldots + X_m}$
Множественный регрессионный анализ	$\underset{\text{(количественная)}}{Y_1} = \underset{\text{(количественные, неколичественные)}}{X_1 + X_2 + \ldots + X_m}$
Совместный анализ	$\underset{\text{(количественные, неколичественные)}}{Y_1 + Y_2 + \ldots + Y_m} = \underset{\text{(неколичественные)}}{X_1 + X_2 + \ldots + X_m}$
Структурное моделирование	$Y_1 = X_{11} + X_{12} + \ldots + X_{1m}$ $Y_2 = X_{21} + X_{22} + \ldots + X_{2m}$ $\ldots$ $\underset{\text{(количественная)}}{Y_k} = \underset{\text{(количественные, неколичественные)}}{X_{k1} + X_{k2} + \ldots + X_{km}}$

МСА обобщает большое число методов и приемов для обработки многомерных статистических данных, которые можно схематически представить в следующем виде (рис. 1.1). Как видно из рисунка, методы подразделяются по признакам числа зависимых переменных, шкал измерения и структуры исследуемой зависимости.

Некоторые из методов проиллюстрированы в табл. 1.1, в которой отражена форма взаимосвязей между данными и показаны уравнения зависимостей, исследуемые с помощью моделей МСА.

На выбор метода существенно влияет форма представления исходной информации (в виде (1.1) или (1.2)), что соответствует измерению переменных в той или иной шкале. Под *шкалой* понимают систему чисел или иных элементов, принятых для оценки или измерения каких-либо величин. Различают номинальные (классификационные), порядковые (ранговые) и количественные (метрические) шкалы.

Номинальная шкала основана на том, что таким характеристикам объектов, как, например, пол, профессия, регион проживания и др., которые невозможно измерить количественно, присваиваются числовые метки, классифицирующие объект по наличию или отсутствию определенного признака. Если, например, признак может быть или не быть у данного объекта, то говорят о переменной с двумя значениями (дихотомическая, бинарная). Так, признак «пол» дает два класса (мужской, женский). Если обозначить один из них нулем, а другой единицей, то можно подсчитывать частоту появления 1 или 0 и проводить дальнейшие статистические процедуры. Если число значений признака больше двух, то он называется категориальным.

Порядковая шкала соответствует более высокому уровню шкалирования. Она предусматривает сопоставление интенсивности определяемого признака у изучаемых объектов (т.е. располагает их по признаку «больше-меньше», но без указания, насколько больше или насколько меньше). Порядковые шкалы широко используются при анализе предпочтений в различных областях экономики, социологии, но, прежде всего, в анализе

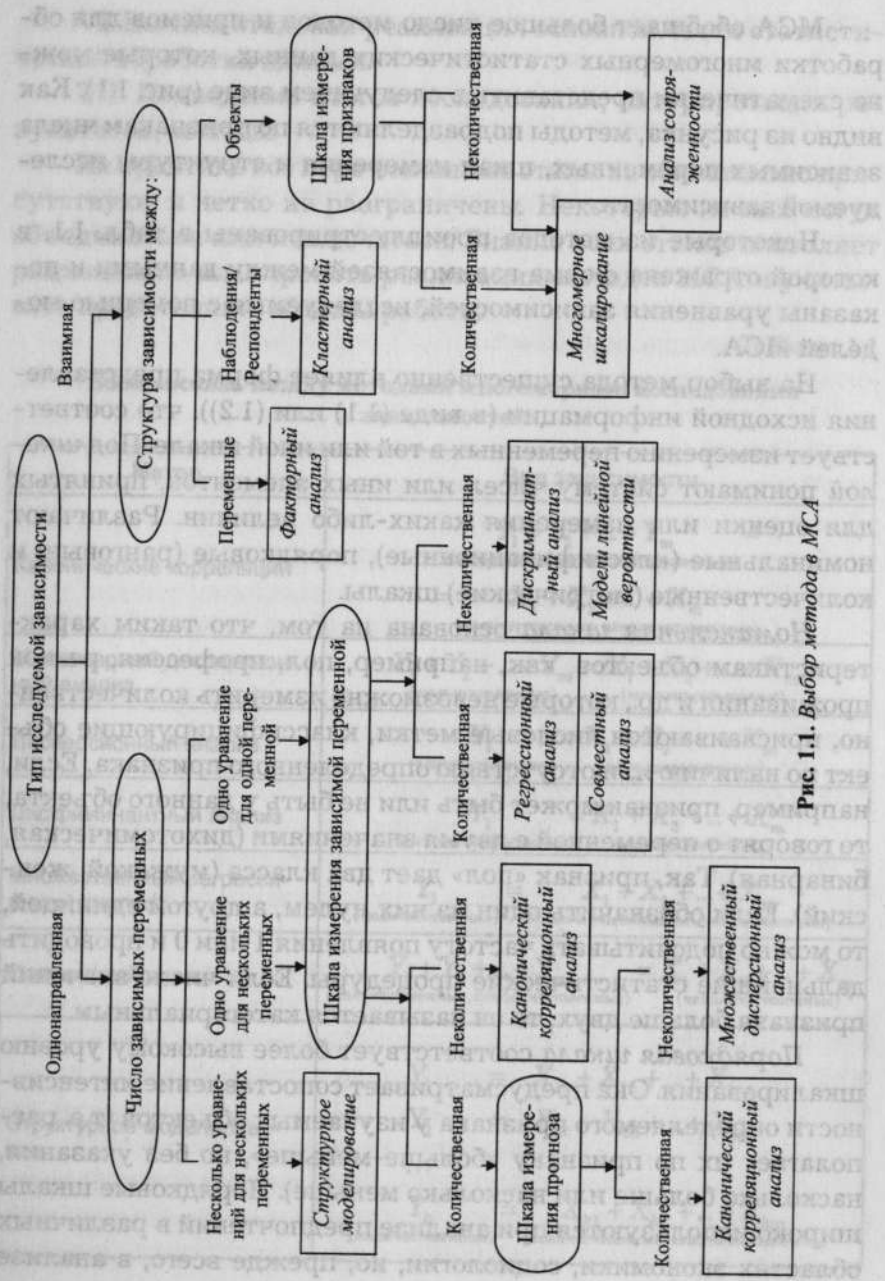


Рис. 1.1. *Выбор метода в МСА*

спроса и потребления. Изучаемые объекты можно обозначить порядковыми числительными (первый, второй, третий), подвергая их любым монотонным преобразованиям (например, возведению в степень, извлечению корня), поскольку первоначальный порядок этим не затрагивается. Порядковую шкалу также называют ранговой, а место объекта в последовательности, которую она собой представляет, рангом объекта. Пример порядковой шкалы — система балльных оценок (школьные оценки, оценки качества продукции и т.д.).

Количественные, или метрические, шкалы подразделяются на два вида: интервальные и пропорциональные. Первые из них, обладая всеми качествами порядковой шкалы, отличаются от нее тем, что точно определяют величину интервала между точками на шкале в принятых единицах измерения. Равновеликость интервалов при этом не требуется. Но она появляется в следующем виде шкал — пропорциональных шкалах. Здесь подразумевается фиксированная нулевая точка отсчета, поэтому пропорциональные шкалы позволяют выяснить, на сколько или во сколько раз один признак объекта больше или меньше другого. С оценками в метрической шкале можно производить различные действия: сложение, умножение, деление. Пример показателя, выраженного в метрической шкале, — объем продукции определенного вида в соответствующих единицах измерения (тонны, рубли).

1.2. Многомерное признаковое пространство

Методы МСА базируются на геометрическом представлении данных. Наблюдаемые объекты располагаются в теоретическом пространстве размерностью, соответствующей числу признаков (элементарных или латентных), которыми они характеризуются. Можно предложить частные случаи признакового пространства: с нулевой размерностью — объекты не имеют характеристик; с единичной размерностью (одномерное

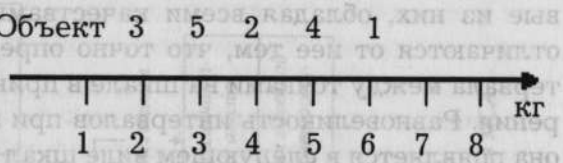
признаковое пространство) — объекты отражаются значениями одного какого-либо признака; многомерное пространство — объекты представлены значениями двух и более признаков (m-мерное признаковое пространство).

Рассмотрим простой пример, когда пять домохозяйств последовательно характеризуются значениями одного, двух и трех признаков.

1. Одномерное признаковое пространство.

Домохозяйство	Средний уровень потребления мяса в месяц на одного человека, кг (X)
1	6
2	4
3	2
4	5
5	3

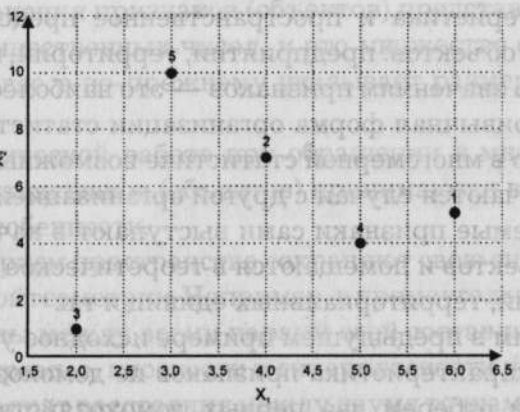
Его можно представить в виде одной градуированной шкалы:



2. Двумерное признаковое пространство.

Домохозяйство	Средний уровень потребления мяса на одного человека, кг (X_1)	Средний уровень потребления молока на одного человека, л (X_2)
1	6	5
2	4	7
3	2	1
4	5	4
5	3	10

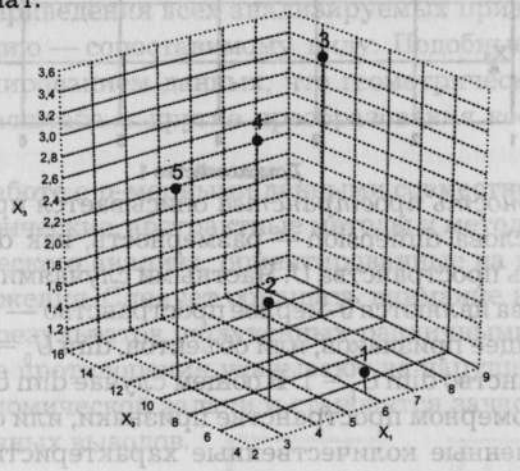
Наблюдаемые объекты геометрически представляются на плоскости в двумерной (декартовой) системе координат.



3. Трехмерное признаковое пространство.

Домохозяйство	Средний уровень потребления мяса на одного человека, кг (X_1)	Среднее потребление молока на одного человека, л (X_2)	Среднее потребление сахара и кондитерских изделий на одного человека, кг (X_3)
1	6	5	1,2
2	4	7	1,9
3	8	12	3,5
4	7	14	2,7
5	3	11	2,8

Наблюдаемые объекты представляются в трехмерной системе координат.



Характеристика и пространственное представление наблюдаемых объектов: предприятий, территорий, групп населения и т. д. по значениям признаков — это наиболее распространенная и привычная форма организации статистических данных. Однако в многомерной статистике возможны и достаточно часто встречаются случаи с другой организацией данных, когда оцениваемые признаки сами выступают в качестве наблюдаемых объектов и помещаются в теоретическое пространство предприятий, территориальных единиц и т.п.

Изменим в предыдущем примере исходное условие: пусть требуется характеристика признаков по домохозяйствам. Для наглядности отберем два первых домохозяйства и покажем возможность размещения их на координатной плоскости признаковых значений:

Размерность пространства описывается при помощи сокращения слова dimension — размерность, как dim U, то есть размерность пространства U. Частными случаями многомерного пространства являются 0-мерное пространство — пространство, не содержащее признаков, или объектов, $\dim U_0 = 0$, и одномерное пространство $\dim U_1 = 1$. В общем случае $\dim U_p = p$.

В многомерном пространстве признаки, или объекты, имеют определенные количественные характеристики. Все принимаемые значения признаков (объектов) представляют собой множества вещественных чисел, и это множество обозначают символом R^p, где p по-прежнему указывает размерность пространства.

В аналитической работе при обращении к многомерному пространству признаков (объектов) принимаются во внимание следующие особенности.

- В p-мерном пространстве сохраняют свою силу положения евклидовой геометрии. Например, в прямоугольной системе координат углы между всеми парами осей составляют 90°, параллельные прямые, плоскости и гиперплоскости не пересекаются, если квадрат расстояния между двумя точками в двумерном пространстве определяется по известной формуле Пифагора: $c^2 = x_1^2 + x_2^2$, то в многомерном пространстве — аналогичным образом $c^2 = x_1^2 + x_2^2 + ... + x_p^2$.
- Пространство, размерность которого превышает три, уже не может быть представлено визуально, и все задачи в этом случае решаются при помощи абстрактной логики и алгебраических методов.
- В многомерном анализе, как правило, используется большое число признаков, разнородных по своей природе. В связи с этим на первом этапе исследований обычно возникает проблема приведения всех анализируемых признаков к одному основанию — сопоставимому виду. Подобные проблемы решают нормированием данных, что геометрически означает изменение масштаба и другие преобразования координатной системы.
- При работе с p-мерными данными совместно используются математические, абстрактные методы и методы экономико-статистического анализа, ориентированные на конкретные сферы приложения. Следует обращать внимание на непротиворечивость результатов, полученных различными методами. Возникающие противоречия указывают на нарушения логики решения экономической задачи и становятся зачастую источником ошибочных выводов.

Вопросы

1. В чем особенности МСА?
2. Основные этапы МСА.
3. Формы представления данных, используемых в МСА.
4. Понятие признакового пространства. Приведите примеры.
5. Виды зависимостей, исследуемых многомерными статистическими методами.

ГЛАВА 2
Многомерная генеральная и выборочная совокупности

Никто не обнимет необъятного.
К. Прутков. Мысли и афоризмы

2.1. Распределение и характеристики многомерной совокупности

Закономерности, которым подчиняется случайная величина, полностью определяются условиями ее наблюдения и математически задаются соответствующими законами распределения вероятностей. Однако при проведении статистических исследований более удобным и распространенным является понятие генеральной совокупности.

Генеральной совокупностью называется множество всех мыслимых наблюдений, которые могли бы быть произведены при данном комплексе условий.

Поскольку в определении идет речь о мыслимо возможных наблюдениях, то генеральная совокупность есть понятие абстрактное и ее не следует смешивать с реальными совокупностями, подлежащими статистическому исследованию. Так, обследовав даже все предприятия отрасли, мы можем рассматривать их как представителей гипотетически возможной более широкой совокупности предприятий, которые могли бы функционировать в рамках комплекса условий.

Генеральная совокупность может быть как конечной, так и бесконечной. Конечная совокупность наблюдается, например, при обследовании школ города, а бесконечная — в научных исследованиях, когда важен ожидаемый результат большого числа экспериментов.

Знание характера распределения случайной величины позволяет исследователю корректно подходить к применению тех или иных методов МСА. В частности, корреляционный анализ применим лишь в том случае, если две случайные величины нормально распределены. Проверка значимости парных и частных коэффициентов корреляции осуществляется также в предпосылке нормальности их распределения.

Теоретически законы распределения можно подразделить на два больших класса: *дискретные* и *непрерывные*. Дискретные распределения отражают прерывность значений случайной величины, среди них наиболее известны биномиальное, гипергеометрическое распределения и распределение Пуассона.

Непрерывные распределения представляют более многочисленный по сравнению с дискретными класс распределений. Это распределения случайных величин, в значениях которых априори нет пропусков. К непрерывным относится и нормальное распределение, которое широко известно и особенно часто встречается и в теоретических разработках и аналитической практике. Теоретическое обоснование тому, что с помощью нормального распределения можно описать подавляющее большинство реально происходящих процессов, дает закон больших чисел. С увеличением числа наблюдений нередко многие другие виды распределений принимают вид нормального. С учетом этих свойств, хорошей разработанности и сравнительно простой формальной структуры нормальное распределение чаще других применяется в многомерной статистике. В классе непрерывных распределений, кроме нормального, известно большое число других видов распределений: гамма-распределение, распределение Стьюдента, распределения Вейбулла, Парето и др.

В зависимости от того, по каким данным строится распределение (по наблюдениям или вычисленным), различают *эмпирическое* и *теоретическое распределения*.

Различают *одномерные* и *многомерные распределения*. Многомерные распределения учитывают значения нескольких признаков одновременно. В простейшем случае генеральная совокупность есть одномерная случайная величина X с функцией распределения $F(x) = P(X < x)$, которая определяется вероятностью того, что X примет значение, меньшее фиксированного действительного числа x.

В МСА изучаются генеральные совокупности с точки зрения нескольких признаков (более 2-х). Рассматриваемое множество признаков обозначается вектором X, имеющим p компонент, каждая из которых характеризует соответствующий признак X_j, $j = 1, 2, ..., p$.

Например, если предметом изучения является множество (пять) предприятий, характеризующихся набором признаков (три), то формально их можно представить в виде матрицы, где строки (векторы) представляют отдельные объекты (наблюдения), а компоненты — векторы признаков, характеризующие объект: $\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ & \cdots & \\ X_{51} & X_{52} & X_{53} \end{bmatrix}$.

В общем виде, если имеется n наблюдений (объектов) над p переменными (признаками), то данные образуют матрицу размерности $n \times p$, i-я строка которой характеризует i-е наблюдение (объект) по всем p признакам:

$$\underset{n \times p}{\mathbf{X}} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix}. \qquad (2.1)$$

Как было показано ранее (в разделе 1.2), данные могут рассматриваться и по столбцам как p признаков, расположенных в n-мерном пространстве объектов.

Под *многомерной генеральной совокупностью* понимаем совокупность, когда на каждом из ее объектов регистрируются значения набора признаков $X_1, ..., X_p$.

Многомерный признак — это p-мерный вектор $\mathbf{X}=(X_1,...,X_p)$ показателей, среди которых могут быть количественные, порядковые и номинальные.

Многомерным наблюдением будем называть результат регистрации значений многомерного признака на i-м статистически обследованном объекте: $\mathbf{X}_i=(X_{i1},...,X_{ip})$, $i=1,...,n$.

Таким образом, исходный массив данных в МСА образуется как последовательность многомерных наблюдений (результат измерения значений p показателей на каждом из n объектов исследуемой совокупности).

Исследуемый многомерный признак интерпретируется как многомерная случайная величина (случайный вектор $\mathbf{X}=(X_1,...,X_p)$ в p-мерном евклидовом пространстве) и соответственно последовательность многомерных наблюдений $\mathbf{X}_i=(X_{i1},...,X_{ip})$, $i=1,...,n$ — как выборка из генеральной совокупности.

Таким образом, по имеющейся случайной выборке $\mathbf{X}_1,...,\mathbf{X}_n$ значений многомерной случайной величины необходимо сделать вывод о ее поведении (свойствах).

Функцией распределения случайного p-мерного вектора $\mathbf{X}$ называется детерминированная неотрицательная величина, определяемая по формуле $F(x_1,x_2,...,x_p)=$ $=P(X_1<x_1,X_2<x_2,...,X_p<x_p)$, где x — p-мерный вектор фиксированных действительных чисел.

Специфика МСА заключается в том, что в отличие от одномерного случая многомерная функция распределения перестает быть исчерпывающей (по информативности) формой задания изучаемого закона распределения. Для описания закона распределения многомерной величины в непрерывном случае* используют *функцию плотности вероятности* $f_x(X_1,X_2,...,X_p)$, которая определяется так, что вероятность попадания значений случайного вектора $\mathbf{X}$ в некоторое заданное подмножество A: $P\{\mathbf{X}\in A\}=\int\limits_A f_x\left(X_1,X_2,...,X_p\right)dX_1dX_2...dX_p$.

Для описания частного закона распределения вероятностей некоторой части компонент $\mathbf{X}=(X_1,...,X_s)$, $s<p$, вектора $\mathbf{X}$ ис-

* Далее рассматриваем именно непрерывный случай.

пользуются частная (маржинальная) функция распределения и частная плотность вероятности, задаваемые соотношениями:

$$F_s(x_1,x_2,...,x_s)=P(X_1<x_1,X_2<x_2,...,X_s<x_s)=$$
$$=P(X_1<x_1,X_2<x_2,...,X_s<x_s,X_{s+1}<\infty,...,X_p<\infty)=$$
$$=F(x_1,x_2,...,x_s,\infty,...,\infty),$$
$$f_s(X_1,X_2,...,X_s)=\int\limits_{X_{s+1}}\int\limits_{X_{s+2}}...\int\limits_{X_p}f_x\left(X_1,...,X_s,X_{s+1},...,X_p\right)dX_{s+1}...dX_p.$$

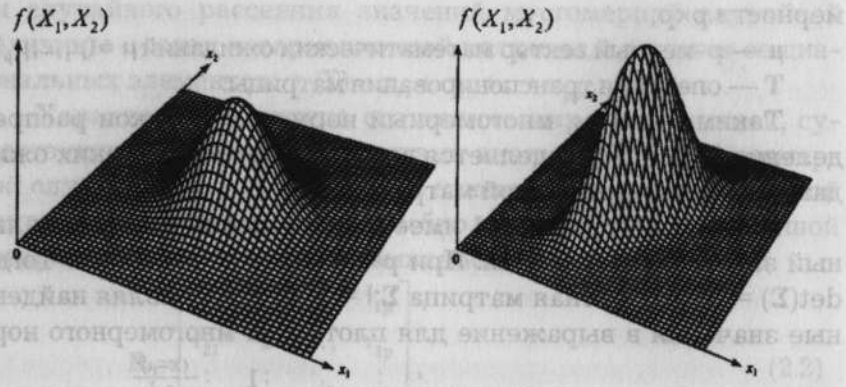
Условная плотность вероятности случайного вектора определяется при условии, что значения другого вектора зафиксированы на некотором уровне. Как и в одномерном случае, для ее определения применяется теорема умножения вероятностей.

На практике для описания многомерной случайной величины, как правило, приходится ограничиваться только информацией, которую дают числовые характеристики закона распределения: вектор средних значений: $\left(E(X_1),E(X_2),...,E(X_p)\right)$ и *матрица ковариаций:* $\Sigma=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \dots & \sigma_{1p}\\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp}\end{bmatrix}$.

Причем математические ожидания $E(X_j)$ определяются по соответствующим одномерным частным плотностям распределения $E(X_j)=\int xf_{X_j}(x)dx$, $j=1,...,p$, а ковариации определяются с использованием двумерных частных плотностей распределения пары $\left(X_j,X_k\right)$, как $\sigma_{jk}=E\left\{\left(X_j-\mu_j\right)\left(X_k-\mu_k\right)\right\}$, где $\mu_j=E(X_j)$, $j=1,...,p$.

Матрица ковариаций Σ характеризует свойства исследуемого многомерного вектора — степень случайного разброса отдельно по каждой компоненте и в целом по многомерному признаку. Диагональные элементы Σ_{ii} матрицы Σ определяют частные дисперсии компонент X_j: $\sigma_{jj}=E\left(X_j-\mu_j\right)^2=V(X_j)$.

Многомерным аналогом дисперсии является величина определителя ковариационной матрицы, называемая *обобщенной дисперсией* многомерной случайной величины $V_{об}(X)=\det(\Sigma)$.



а) б)

Рис. 2.1. *Двумерное нормальное распределение при* $\sigma_{11} = \sigma_{22}$ *и а)* $r_{X_1X_2} = 0{,}00$ *б)* $r_{X_1X_2} = 0{,}75$.

2.3. Статистические оценки многомерной генеральной совокупности

Задачи математической статистики фактически сводятся к обоснованному суждению об объективных свойствах генеральной совокупности по результатам выборки. *Выборкой* из генеральной совокупности называют результаты ограниченного ряда наблюдений $X_1, X_2, ..., X_n$, где n — объем выборки. Достоверность выводов, получаемых в результате статистической обработки данных, во многом зависит от успешного решения вопроса представительности выборки — полноты и адекватности представления в выборке свойств анализируемой генеральной совокупности.

Необходимость выборочного обследования при решении практических задач связана со следующими причинами:

1) генеральная совокупность настолько многочисленна, что исследование всех ее элементов слишком трудоемко. С такой ситуацией приходится встречаться, например, при

контроле качества продукции крупносерийного и массового производства;

2) бесконечно большая генеральная совокупность, когда даже значительное множество наблюдений не исчерпывает всей совокупности. Например, при разработке статистически обоснованных временных нормативов на изготовление изделия. В этом случае выводы, полученные по результатам конечного числа наблюдений, распространяются на всю генеральную совокупность, охватывающую все детали, которые могут быть изготовлены на оборудовании данного типа на протяжении ряда последующих лет;

3) когда в процессе проведения испытания происходит разрушение отбираемых образцов (например, испытание предела прочности).

Точечные оценки параметров многомерной генеральной совокупности.

Выборку объема n из p-мерной генеральной совокупности X можно представить в виде матрицы данных вида (2.1), строки которой рассматриваются как n независимых реализаций p-мерного случайного вектора X с плотностью распределения $f_x(X_1, X_2, ..., X_p)$. Таким образом, элементы X_{ij} матрицы $\mathbf{X}$ можно рассматривать либо как случайные (одномерные) величины, независимые по i, либо как конкретные наблюдаемые значения — координаты n точек в p-мерном евклидовом пространстве.

Приведем точечные оценки моментов генеральной совокупности, которые получили наибольшее практическое применение.

Оценка начального момента m-го порядка l-й компоненты случайного вектора X вычисляется по формуле

$$\overline{x_l^m} = \frac{1}{n}\sum_{i=1}^{n} X_{il}^m, \; l = 1, ..., p.$$

При $m = 1$ получим оценку математического ожидания (среднюю арифметическую): $E(X_l) = \overline{X}_l = \frac{1}{n}\sum_{i=1}^{n} X_{il}, \; l = 1, ..., p.$

Оценка ковариационной матрицы $\sum$ (матрицы выборочных дисперсий и коэффициентов ковариации) случайного вектора

X определяется как $\mathbf{S}=\begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{bmatrix}$, где недиагональные элементы — выборочные коэффициенты ковариации l-й и j-й компонент вектора X: $s_{lj}=\frac{1}{n}\sum_{i=1}^{n}\left(X_{il}-\bar{X}_l\right)\left(X_{ij}-\bar{X}_j\right)$ $l,j=1,\dots,p; l\neq j$

Причем диагональные элементы s_{ll} представляют собой выборочные дисперсии l-й компоненты случайного вектора X:

$$s_{ll}=V(X_l)=\frac{1}{n}\sum_{i=1}^{n}\left(X_{il}-\bar{X}_l\right)^2, l=1,\dots,p.$$

Вместо $\mathbf{S}$ употребляют также несмещенную оценку матрицы $\sum$: $\hat{\mathbf{S}}=\frac{n}{n-1}\mathbf{S}$.

Могут быть определены среднеквадратические (стандартные) отклонения элементов вектора X: $s_l=\sqrt{s_{ll}}, l=1,\dots,p$.

Оценки элементов корреляционной матрицы (2.2) можно получить по формуле

$$r_{lj}=\frac{s_{lj}}{\sqrt{s_{ll}s_{jj}}}=\frac{\sum_{i=1}^{n}\left(X_{il}-\bar{X}_l\right)\left(X_{ij}-\bar{X}_j\right)}{\sqrt{\sum_{i=1}^{n}\left(X_{il}-\bar{X}_l\right)^2}\sqrt{\sum_{i=1}^{n}\left(X_{ij}-\bar{X}_j\right)^2}}, l,j=1,\dots,p; l\neq j,$$

где r_{lj} — оценка парного коэффициента корреляции между l-й и j-й компонентами вектора X.

Часто от исходных выборочных значений переходят к так называемым стандартизованным, выполняя центрирование (вычитание математического ожидания) и нормирования (деления на стандартное отклонение):

$$\mathbf{Z}=\begin{bmatrix} Z_{11} & \dots & Z_{1p} \\ \vdots & \ddots & \vdots \\ Z_{n1} & \dots & Z_{np} \end{bmatrix}, \text{где } Z_{ij}=\frac{X_{ij}-\bar{X}_j}{s_j}, i=1,\dots,n, j=1,\dots,p. \quad (2.3)$$

Тогда корреляционная матрица имеет представление $\mathbf{R}=\frac{1}{n}\mathbf{Z}^T\mathbf{Z}$. Отметим, что стандартизованные переменные имеют нулевое математическое ожидание и единичную дисперсию: $E(Z_j)=0,\ V(Z_j)=1, j=1,2,\dots,p$.

Доверительные области. При малом объеме выборки точечные оценки могут достаточно далеко отклоняться от оцениваемых параметров и поэтому вводится понятие интервальной оценки параметра генеральной совокупности.

Определения и понятия интервального оценивания можно перенести на случай векторного параметра $\Theta=\left(\theta_1,\theta_2,\dots,\theta_k\right)^T$ с заменой доверительного интервала доверительной областью в соответствующем p-мерном пространстве.

Доверительной областью вектора параметров Θ генеральной совокупности называется случайная область, полностью определяемая результатами наблюдений, которая с близкой к единице доверительной вероятностью (надежностью) γ содержит неизвестное значение вектора Θ. Стремятся определить доверительные области, имеющие минимальные размеры при данной надежности γ. Часто этому условию удовлетворяют области, симметричные относительно вектора оценок $\hat{\Theta}$ параметров Θ.

Основную трудность в построении доверительной области представляет определение законов распределения подходящих статистик. В настоящее время эти вопросы хорошо разработаны только для нормального распределения наблюдаемых случайных величин.

Доверительная область для вектора математических ожиданий. Пусть по результатам n наблюдений из генеральной совокупности X с p-мерным нормальным распределением $N_p(\boldsymbol{\mu},\sum)$ найдены вектор средних $\bar{\mathbf{X}}=\left(\bar{X}_1,\bar{X}_1,\dots,\bar{X}_p\right)$ и несмещенная оценка матрицы ковариаций $\sum$. Требуется найти с надежностью γ (или уровнем значимости $\alpha=1-\gamma$) доверительную область для p-мерного вектора генеральных средних $\boldsymbol{\mu}$.

Предположим, что ковариационная матрица известна. Напомним, что для одномерной нормально распределенной гене-

ральной совокупности доверительный интервал для μ находится из формул $|t| < \Phi^{-1}(\gamma), t = \frac{\bar{x}-\mu}{\sigma}\sqrt{n}$, где Φ — функция Лапласа, а статистика t подчиняется стандартному нормальному закону $N(0, 1)$.

Последнее равенство можно переписать в виде $t^2 = n(\bar{x}-\mu)(\sigma^2)^{-1}(\bar{x}-\mu)$ и обобщить на случай p-мерной совокупности следующим образом:

$$n(\bar{\mathbf{X}}-\boldsymbol{\mu})^T \mathbf{S}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu}) \le \chi^2_{1-\gamma}(p), \qquad (2.4)$$

где $\chi^2(p)$ — квантиль χ^2-распределения для числа степеней свободы p.

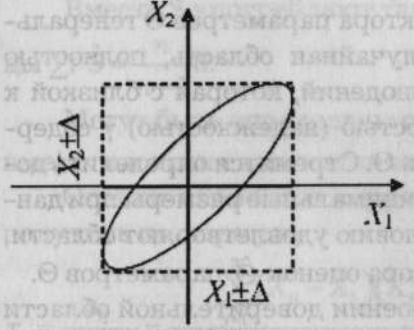


Рис. 2.2. *Доверительная область для средних значений двух признаков*

Формула (2.4) задает доверительную область для вектора математических ожиданий многомерной нормально распределенной случайной величины. Пусть теперь ковариационная матрица неизвестна. Чтобы при $p=1$ построить доверительный интервал для μ, используют статистику $t = \frac{\bar{x}-\mu}{s}\sqrt{n}$, которая имеет t-распределение Стьюдента с $n-1$ степенями свободы.

Равенство можно переписать в эквивалентной форме $t^2 = n(\bar{x}-\mu)(s^2)^{-1}(\bar{x}-\mu)$.

По аналогии строится T^2 статистика Хотеллинга, которую используют при построении доверительной области для вектора средних p-мерной генеральной совокупности:

$$T^2 = n(\bar{\mathbf{X}}-\boldsymbol{\mu})^T \mathbf{S}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu}),$$

где $\mathbf{S}^{-1}$ — матрица, обратная матрице несмещенных оценок $\mathbf{S}$.

Учитывая, что распределения Фишера F и Хотеллинга T^2 связаны соотношением $T^2_{\alpha;(p,n-p)} = \frac{p(n-1)}{n-p}F_{\alpha;(p,n-p)}$, получим уравнение поверхности, ограничивающей доверительную область для вектора средних с надежностью $\gamma=1-\alpha$:

$$n(\bar{\mathbf{X}}-\boldsymbol{\mu})^T \mathbf{S}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu}) = \frac{p(n-1)}{n-p}F_{\alpha;(p,n-p)},$$

где $F_{\alpha;(p,n-p)}$ — точка F-распределения, соответствующая уровню значимости α и числам степеней свободы p и $n-p$.

Последнее уравнение определяет p-мерный эллипсоид (эллипс при $p=2$, см. рис. 2.2) с центром $\bar{\mathbf{X}}$, так как его левая часть представляет собой положительно определенную квадратичную форму относительно $\boldsymbol{\mu}$.

2.4. Проверка статистических гипотез о параметрах многомерной нормально распределенной генеральной совокупности

Проверка гипотезы о равенстве вектора средних значений постоянному вектору.

Рассмотрим p-мерную генеральную совокупность с распределением $N_p(\boldsymbol{\mu}, \Sigma)$, $\det(\Sigma)\neq 1$. По выборке объема n определим вектор средних арифметических и несмещенную оценку ковариационной матрицы.

Если ковариационная матрица известна, то для проверки гипотезы о равенстве вектора генеральных средних заданному значению H_0: $\boldsymbol{\mu}=\boldsymbol{\mu}_0$ против альтернативы H_1: $\boldsymbol{\mu}\neq\boldsymbol{\mu}_0$ употребляется статистика $\chi^2 = n(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)^T \Sigma^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)$, имеющая χ^2 распределение Пирсона с числом степеней свободы p при справедливости гипотезы H_0.

Если же ковариационная матрица Σ неизвестна, то можно воспользоваться статистикой Хотеллинга:

$$T^2 = n(\bar{X}-\boldsymbol{\mu}_0)^T S^{-1}(\bar{X}-\boldsymbol{\mu}_0).$$

Критическую область, определяемую с учетом $T^2_{\alpha;(p,n-p)} = \frac{p(n-1)}{n-p} F_{\alpha;(p,n-p)}$, можно вычислить с помощью таблиц F-распределения Фишера.

Сравнение двух генеральных совокупностей.

Будем называть две генеральные совокупности однородными, если, кроме одних и тех же признаков, они имеют одинаковые законы распределения вероятностей.

Рассмотрим две нормально распределенные совокупности X и Y. Их распределения полностью задаются параметрами μ_X, $\sum_X$ и μ_Y, $\sum_Y$. Следовательно, для проверки однородности совокупностей достаточно сравнить их ковариационные матрицы $\sum_X$ и $\sum_Y$. Затем, в случае принятия гипотезы о равенстве этих ковариационных матриц, сравнить генеральные средние совокупностей μ_X и μ_Y.

Для сравнения матриц коэффициентов ковариации проверяется гипотеза $H_0: \sum_X = \sum_Y$ против альтернативы $H_1: \sum_X \neq \sum_Y$ на основе выборок из совокупностей соответственно объемов n_X и n_Y.

В качестве значения статистики Бартлетта берется случайная величина $W = ba$, где

$$b = 1 - \left(\frac{1}{n_X - 1} + \frac{1}{n_Y - 1} - \frac{1}{n_X + n_Y - 2} \right) \frac{2p^2 + 3p - 1}{6(p+1)},$$

$$a = (n_X + n_Y - 2) \ln \det \mathbf{S}_{XY} - \left[(n_X - 1) \ln \det \mathbf{S}_X + (n_Y - 1) \ln \det \mathbf{S}_Y \right],$$

$\mathbf{S}_X$ — несмещенная оценка ковариационной матрицы для совокупности X,

$\mathbf{S}_Y$ — несмещенная оценка ковариационной матрицы для совокупности Y.

$\mathbf{S}_{XY} = \frac{1}{n_X + n_Y - 2} \left[(n_X - 1)\mathbf{S}_X + (n_Y - 1)\mathbf{S}_Y \right]$ – несмещенная оценка объединенной ковариационной матрицы.

При справедливости гипотезы H_0, достаточно больших объемах выборок n_X и n_Y и достаточно малой величине

$$c = \frac{p(p+1)}{48b^2} \left\{ (p-1)(p+2) \left[\frac{1}{(n_X - 1)^2} + \frac{1}{(n_Y - 1)^2} - \frac{1}{(n_X - 1)^2 + (n_Y - 1)^2} \right] - 6(1-b)^2 \right\}$$

статистика W аппроксимируется χ^2 распределением с числом степеней свободы $\frac{p(p+1)}{2}$.

Таким образом, критическая область имеет вид:

$$W > W_{кр} = \chi^2_{\alpha;\frac{p(p+1)}{2}}.$$

Если $W_{набл.} = ba$ попадает в критическую область ($W_{набл} > W_{кр}$), то гипотеза H_0 отклоняется с вероятностью ошибки α, и считается доказанным, что ковариационные матрицы $\sum_X$ и $\sum_Y$ неодинаковы, а следовательно, генеральные совокупности неоднородны.

Иначе, если H_0 не может быть отклонена, то ковариационные матрицы одинаковы. При таком условии необходимо сравнить генеральные средние, т. е. проверить гипотезу H_0: $\mu_X = \mu_Y$ против альтернативы H_1: $\mu_X \neq \mu_Y$.

Для проверки применяется статистика Хотеллинга:

$$T^2 = \frac{n_X n_Y}{n_X + n_Y} \left(\bar{\mathbf{X}} - \bar{\mathbf{Y}} \right)^T \mathbf{S}^{-1}_{XY} \left(\bar{\mathbf{X}} - \bar{\mathbf{Y}} \right).$$

Если гипотеза справедлива, то статистики T^2 и F связаны формулой $T^2_{кр} = \frac{p(n_X + n_Y - 2)}{n_X + n_Y - p - 1} F_{\alpha;(p;n_X+n_Y-p-1)}$, где $F_{\alpha;(p;n_X+n_Y-p-1)}$ находится по таблицам распределения Фишера.

Критическая область имеет вид $T^2 > T^2_{кр}$. Если гипотеза H_0 отвергается с вероятностью ошибки α, то считается доказанной неоднородность генеральных совокупностей X и Y. Иначе считаем, что генеральные совокупности однородны с надежностью $\gamma = 1 - \alpha$.

2.5. Моделирование значений случайных векторов

Решение многих прикладных задач, в частности, проведение модельных (машинных) экспериментов с помощью математических методов, имитационные схемы, требует моделирования случайных векторов.

Исходными данными в такой задаче, как правило, являются характеристики моделируемого случайного p-мерного вектора $\xi=(\xi_1,\ldots,\xi_p)^T$: матрица ковариаций $\mathbf{K}_\xi=\begin{bmatrix} k_{11} & k_{12} & \cdots & k_{1p} \\ k_{21} & k_{22} & \cdots & k_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ k_{p1} & k_{p2} & \cdots & k_{pp} \end{bmatrix}$, где $k_{ij}=k_{ji}=\operatorname{cov}(\xi_i,\xi_j)=E(\xi_i\xi_j)$ и вектор математических ожиданий признаков (составляющих случайного вектора) $E(\xi)=(E(\xi_1),\ldots,E(\xi_p))^T$, причем предполагается одинаковый (одномерный) для всех признаков закон распределения.

Пусть имеется последовательность некоррелированных случайных величин $u_i, i=1,\ldots,p$, имеющих нормальное распределение с нулевым математическим ожиданием и единичной дисперсией, которая без труда получается в большинстве пакетов прикладных программ с помощью датчика псевдослучайных чисел (если $\zeta_1,\ldots,\zeta_{12},\ldots,\zeta_N$ — равномерно распределенные на отрезке [0, 1] псевдослучайные числа, то $u_i=\sum_{h=12(i-1)+1}^{12i}\zeta_h$ имеют стандартное нормальное распределение).

Координаты выходного вектора ξ: $\xi_1,\xi_2,\ldots\xi_p$ могут быть получены по значениям нормально распределенных независимых случайных величин $u_1,u_2,\ldots,u_p$ следующим образом:

$\xi_i=a_{i1}u_1+a_{i2}u_2+\cdots+a_{ii}u_i+E(\xi_i)$, $i=1,\ldots,p$ или

$$\begin{cases} \xi_1=a_{11}u_1+E(\xi_1), \\ \xi_2=a_{21}u_1+a_{22}u_2+E(\xi_2), \\ \xi_3=a_{31}u_1+a_{32}u_2+a_{33}u_3+E(\xi_3), \\ \ldots \\ \xi_p=a_{p1}u_1+a_{p2}u_2+\cdots+a_{pp}u_p+E(\xi_p). \end{cases}$$

Можно переписать систему линейных уравнений в матричном виде: $\xi=\mathbf{A}\mathbf{Ч}\mathbf{U}+E(\xi)$, где $\mathbf{A}=\begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix}$, $\mathbf{U}=\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{bmatrix}$.

Элементы матрицы $\mathbf{A}$ подлежат определению. Выразим их через элементы матриц $E(\xi),\mathbf{K}_\xi$.

Так как $\tilde{\xi}=\xi_i-E(\xi_i)$, то $\xi_i=\tilde{\xi}_i+E(\xi_i)$, поэтому будем рассматривать центрированные случайные величины, прибавив к которым соответствующие математические ожидания, получим искомые координаты вектора.

Рассмотрим ковариацию двух случайных величин $\tilde{\xi}_i,\tilde{\xi}_j$: $k_{ij}=E(\tilde{\xi}_i\tilde{\xi}_j)$. Так как $\tilde{\xi}_i=a_{i1}u_1+a_{i2}u_2+\cdots+a_{ii}u_i=\sum_{l=1}^{i}a_{il}u_l$, аналогично $\tilde{\xi}_j=\sum_{m=1}^{j}a_{jm}u_m$, то с учетом свойств математического ожидания: $k_{ij}=E\left(\sum_{l=1}^{i}a_{il}u_l\sum_{m=1}^{j}a_{jm}u_m\right)=E\left(\sum_{l=1}^{i}\sum_{m=1}^{j}a_{il}a_{jm}(u_lu_m)\right)=$

$$=\sum_{l=1}^{i}\sum_{m=1}^{j}a_{il}a_{jm}E(u_lu_m)=\sum_{l=1}^{i}\sum_{m=1}^{j}a_{il}a_{jm}\operatorname{cov}(u_l,u_m),$$

так как $\begin{cases} \operatorname{cov}(u_l,u_m)=0, & l\neq m, \\ \operatorname{cov}(u_l,u_m)=1, & l=m, \end{cases}$ то между элементами ковариационной матрицы $\mathbf{K}_\xi$ и элементами матрицы линейного преобразования $\mathbf{A}$ имеется следующая связь $k_{ij}=\sum_{k=1}^{\min(i,j)}a_{ik}a_{jk}$, или в матричном виде $\mathbf{K}_\xi=\mathbf{A}\mathbf{A}^T$.

Так как $\mathbf{A}$ нижнетреугольная матрица $(j\le i)$ и $k_{ij}=k_{ji}$, то $k_{ij}=\sum_{k=1}^{j}a_{ik}a_{jk}$ и $k_{ij}=\sum_{k=1}^{j-1}a_{ik}a_{jk}+a_{ij}a_{jj}$, так что:

$$\begin{cases} a_{ij}=\dfrac{k_{ij}-\sum_{k=1}^{j-1}a_{ik}a_{jk}}{a_{jj}}, & j<i \\ & i=1,\ldots,p, j=1,\ldots,i. \\ a_{jj}=\sqrt{k_{jj}-\sum_{k=1}^{j-1}a_{jk}^2}, & j=i, \end{cases}$$

Эти рекуррентные соотношения позволяют найти элементы матрицы $\mathbf{A}$.

Например, пусть $p=2$, $E(\xi)=(1,1)^T$ и $\mathbf{K}_\xi=\begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$.

Тогда последовательно получаем $a_{11}=\sqrt{k_{11}}=\sqrt{4}=2$,

$a_{21}=\dfrac{k_{21}}{a_{11}}=2/2=1$, $a_{22}=\sqrt{k_{22}-a_{21}^2}=\sqrt{3-1}=1{,}41$.

И, следовательно, $\xi_{1l}=a_{11}u_1+E(\xi_1)$, $\xi_{2l}=a_{21}u_1+a_{22}u_2+E(\xi_2)$ или $\xi_{1l}=2u_1+1, \xi_{2l}=u_1+1{,}41u_2+1$. Если имеются значения стандартного нормального распределения u: -0,62; -0,42; 0,557; 1,112; -0,83; 1,203; -0,61; 0,189; -0,87; -0,76; -0,07; 1,358; 1,338; 0,324; -2,1; 0,617; -0,48; 1,997; -1,26; -0,68, тогда получаем $\xi_{11}=2\cdot(-0{,}62)+1=0{,}24, \xi_{21}=-0{,}62+1{,}41\cdot(-0{,}42)+1=-0{,}22$, т.е. первое значение случайного вектора $\xi_1=(0{,}24,-0{,}22)$, далее $\xi_{12}=2\cdot 0{,}557+1=2{,}11; \xi_{21}=0{,}557+1{,}41\cdot 1{,}112+1=3{,}13$ и $\xi_2=(2{,}11\ 3{,}13)$ и т.п. В результате получим 10 значений случайного вектора ξ (по столбцам таблицы):

1	2	3	4	5	6	7	8	9	10
-0,24	2,114	-0,66	-0,22	-0,74	0,868	3,677	-3,2	0,044	-1,52
-0,22	3,13	1,872	0,655	-0,95	2,854	2,797	-0,23	3,346	-1,22

Вычислим по полученным значениям случайного вектора математические ожидания признаков и ковариационную матрицу. Получим $E(\hat{\xi})=\begin{pmatrix}0{,}68 & 1{,}45\end{pmatrix}^T$ и $\hat{\mathbf{K}}_\xi=\begin{bmatrix}3{,}59 & 2{,}13\\ 2{,}13 & 3{,}21\end{bmatrix}$. Видно, что векторы математических ожиданий $E(\xi)$ и $E(\hat{\xi})$, а также ковариационные матрицы $\mathbf{K}_\xi$ и $\hat{\mathbf{K}}_\xi$ не совпадают по причине малого объема полученной выборочной совокупности (10 наблюдений).

Пример. 2.1. Данные об издержках на транспортировку продуктов питания 10 фирм, занимающихся снабжением, представлены в таблице. Необходимо построить 95% доверительные интервалы для средних значений трех имеющихся признаков в предположении, что они имеют нормальное распределение, а также доверительную область для первых двух признаков.

Решение.

Вычислим вектор средних значений и ковариационную матрицу для трехмерного нормального распределения.

№ п.п.	Затраты топлива, л, X_1	Затраты на ремонт, у.е., X_2	Капитал фирмы, тыс у.е., X_3
1	16,44	12,43	11,23
2	7,19	2,70	3,92
3	9,92	1,35	9,75
4	4,24	5,78	7,78
5	11,20	5,05	10,67
6	14,25	5,78	9,88
7	13,50	10,98	10,60
8	13,32	14,27	9,45
9	29,11	15,09	3,28
10	12,68	7,61	10,23

Применяя формулу $\bar{X}_l=\dfrac{1}{n}\sum_{i=1}^{n}X_{il}, l=1,2,3$, получим, например, для X_1: $\bar{X}_1=(16{,}44+7{,}19+9{,}92+4{,}24+\ldots+12{,}68)/10=13{,}185$. Аналогично получаем $\bar{X}_2=8{,}104$, $\bar{X}_3=8{,}679$, и вектор средних имеет вид: $\bar{\mathbf{X}}=[13{,}185,\ 8{,}104,\ 8{,}679]$.

Элементы ковариационной матрицы вычисляем по формуле $s_{lj}=\dfrac{1}{10}\sum_{i=1}^{10}\left(X_{il}-\bar{X}_l\right)\left(X_{ij}-\bar{X}_j\right) l,j=1,2,3$. Например, для $i=j=1$: $s_{11}=[(16{,}44-13{,}185)^2+(7{,}19-13{,}185)^2+\ldots+(12{,}68-13{,}185)^2]/10=44{,}029$ и далее, для $i=1$, $j=2$: $s_{12}=[(16{,}44-13{,}185)(12{,}43-8{,}104)+$ $+(7{,}19-13{,}185)(2{,}7-8{,}104)+\ldots+(12{,}68-13{,}185)(7{,}61-8{,}104]/10=20{,}615$ и т.д. Ковариационная матрица имеет вид:

$\mathbf{S}=\begin{bmatrix}44{,}029 & 20{,}615 & -4{,}735\\ 20{,}615 & 23{,}222 & -0{,}547\\ -4{,}735 & -0{,}547 & 8{,}039\end{bmatrix}$ или несмещенная оценка ковариационной матрицы:

$$\hat{\mathbf{S}}=\frac{10}{10-1}\begin{bmatrix}44{,}029 & 20{,}615 & -4{,}735\\ 20{,}615 & 23{,}222 & -0{,}547\\ -4{,}735 & -0{,}547 & 8{,}039\end{bmatrix}=\begin{bmatrix}48{,}921 & 22{,}905 & -5{,}261\\ 22{,}905 & 25{,}803 & -0{,}608\\ -5{,}261 & -0{,}608 & 8{,}933\end{bmatrix}.$$

Вычислим

$$T^2_{\alpha;(p,n-p)}=\frac{p(n-1)}{n-p}F_{\alpha;(p,n-p)}=\frac{3(10-1)}{10-3}F_{0{,}05;(3,7)}=3{,}857\cdot 4{,}347=16{,}767.$$

Тогда получим доверительные интервалы для каждой из средних:

$$13{,}185 - \sqrt{16{,}767}\sqrt{\frac{48{,}921}{10}} \le \mu_1 \le 13{,}185 + \sqrt{16{,}767}\sqrt{\frac{48{,}921}{10}} \text{ и}$$

$4{,}128 \le \mu_1 \le 22{,}242,$

$$8{,}104 - \sqrt{16{,}767}\sqrt{\frac{25{,}803}{10}} \le \mu_2 \le 8{,}104 + \sqrt{16{,}767}\sqrt{\frac{25{,}803}{10}} \text{ и}$$

$1{,}527 \le \mu_2 \le 14{,}681,$

$$8{,}679 - \sqrt{16{,}767}\sqrt{\frac{8{,}933}{10}} \le \mu_3 \le 8{,}679 + \sqrt{16{,}767}\sqrt{\frac{8{,}933}{10}} \text{ и}$$

$4{,}809 \le \mu_3 \le 12{,}549.$

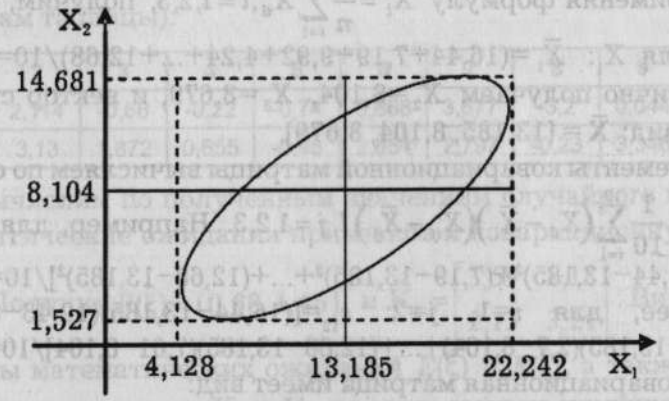


Рис. 2.3. *Доверительная область для средних значений первых двух признаков*

Для первых двух признаков доверительная область строится с учетом формулы $n(\bar{X}-\boldsymbol{\mu}_0)^T \mathbf{S}^{-1}(\bar{X}-\boldsymbol{\mu}_0) = \frac{p(n-1)}{n-p} F_{\alpha;(p,n-p)}$. Имеем:

$$10\left[13{,}185-\mu_1,\ 8{,}104-\mu_2\right]^T \begin{bmatrix} 48{,}921 & 22{,}905 \\ 22{,}905 & 25{,}803 \end{bmatrix}^{-1} \begin{bmatrix} 13{,}185-\mu_1 \\ 8{,}104-\mu_2 \end{bmatrix} =$$

$$= 0{,}35(13{,}185-\mu_1)^2 + 0{,}66(8{,}104-\mu_2)^2 - 2\cdot 0{,}31(13{,}185-\mu_1)(8{,}104-\mu_2) \le$$

$\le 16{,}767.$

Получившаяся доверительная область показана на рис. 2.3 в виде эллипса.

Пример 2.2. С целью оценки воздействия состояния окружающей среды на здоровье населения собраны данные* по двум федеральным округам. Необходимо проверить при $\alpha = 0{,}05$ существенность различий двух округов по выбранным двум показателям.

№ региона	Северо-Западный федеральн. округ		Центральный федеральн. округ	
	Число умерших на 1000 чел. населения	Заболеваемость на 1000 чел. населения новообразованиями	Число умерших на 1000 чел. населения	Заболеваемость на 1000 чел. населения новообразованиями
1	16,6	9,6	16,1	12,3
2	12,5	7,6	17,6	8,5
3	15,3	8,1	19,2	10
4	17,1	7,6	18,2	8
5	16,3	7,3	20,2	10,9
6	20,1	6,4	18,1	7,5
7	11,6	10,3	19,2	8,2
8	20,9	8,7	18,2	6,8
9	22,5	7,4	17	10
10	16,4	8,6	18,1	9,3
11			17,7	10,9
12			19,7	9,8
13			19,9	6,7
14			18,2	8,9
15			21,9	8
16			21,5	9,2
17			19,5	11,3
18			15,6	10,3
$\bar{X}$	16,93	8,16	18,66	9,26

Решение.

1. Определим векторы средних и ковариационные матрицы.

$\bar{X}_1 = [16{,}93,\ 8{,}16];\ \bar{X}_2 = [18{,}66,\ 9{,}26];$

$$\mathbf{S}_1 = \begin{bmatrix} 13{,}38 & -1{,}86 \\ -1{,}86 & 1{,}50 \end{bmatrix},\ \mathbf{S}_2 = \begin{bmatrix} 2{,}93 & -0{,}76 \\ -0{,}76 & 2{,}62 \end{bmatrix}.$$

*Данные за 2001 г. по: Регионы России. — М.: Госкомстат, 2002.

Объединенная ковариационная матрица рассчитывается так: $\mathbf{S}_{12} = \frac{1}{10+18-2}((10-1)\mathbf{S}_1 + (18-1)\mathbf{S}_2) = \begin{bmatrix} 6{,}54 & -1{,}14 \\ -1{,}14 & 2{,}24 \end{bmatrix}$.

Найдем обратную матрицу $\mathbf{S}_{12}^{-1} = \begin{bmatrix} 0{,}17 & 0{,}09 \\ 0{,}09 & 0{,}49 \end{bmatrix}$.

2. Рассчитаем фактическое значение критерия Хоттелинга:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right)^T \mathbf{S}_{12}^{-1}\left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right) =$$

$$= \frac{10 \cdot 18}{10+18}[16{,}93-18{,}66,\ \ 8{,}16-9{,}26]^T \begin{bmatrix} 0{,}17 & 0{,}09 \\ 0{,}09 & 0{,}49 \end{bmatrix} [16{,}93-18{,}66,\ \ 8{,}16-9{,}26].$$

3. Найдем критическое значение критерия Хоттелинга и сравним с фактическим

$$T^2_{\text{кр}} = \frac{p(n_X + n_Y - 2)}{n_X + n_Y - p - 1} F_{\alpha;(p;n_X+n_Y-p-1)} = \frac{2(10+18-2)}{10+18-2-1} F_{0{,}05;(2,25)} = 7{,}04.$$

Поскольку критическое значение больше фактического, то гипотеза о равенстве векторов средних значений признаков для двух округов не может быть отвергнута.

4. Проверим с помощью критерия Бартлетта равенство ковариационных матриц двух выборок.

Рассчитаем $b = 1 - \left(\frac{1}{n_X - 1} + \frac{1}{n_Y - 1} - \frac{1}{n_X + n_Y - 2}\right)\frac{2p^2 + 3p - 1}{6(p+1)} =$

$=0{,}905,$

$$a = (n_X + n_Y - 2)\ln\det\mathbf{S}_{XY} - \left[(n_X - 1)\ln\det\mathbf{S}_X + (n_Y - 1)\ln\det\mathbf{S}_Y\right] =$$

$=8{,}835.$

$W = ab = 7{,}99$. По таблице находим $\chi^2_{0{,}05;3} = 7{,}81$. Таким образом, фактическое значение критерия Бартлетта больше табличного и гипотеза о равенстве ковариационных матриц и, следовательно, однородности двух федеральных округов по выбранным двум признакам отвергается с надежностью 95%.

Вопросы и задачи

1. Имеются данные о нескольких индивидах:

№ п.п.	Число лет образования	Возраст	Логарифм доходов, руб./месяц
1	10	67	6,55
2	15	23	8,16
3	15	42	7,24
4	10	60	6,40
5	13	29	7,90
6	10	27	9,29
7	10	59	6,68
8	17	30	7,31
9	13	20	0,00
10	13	55	7,78
11	18	60	8,07

Вычислите вектор средних значений и ковариационную матрицу. Прокомментируйте результаты.

2. Для данных предыдущей задачи проверьте гипотезу о равенстве вектора средних, вычисленного по первым 5 наблюдениям, вектору средних, вычисленного по последним 6 наблюдениям.

3. Совместная плотность двумерной случайной величины (X_1, X_2) задана формулой $f(X_1, X_2) = \frac{1}{1{,}6\pi} e^{-\frac{1}{1{,}28}[(X_1-2)^2 - 1{,}2(X_1-2)(X_2+3)+(X_2+3)^2]}$.

Найти $\bar{X}_1$, $\bar{X}_2$, σ_{11}, σ_{22} и r_{12}.

4. Пусть $\mathbf{X}$ имеет нормальное распределение $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ с $\boldsymbol{\mu} = [-3, 1, 4]$ и $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$. Какие из следующих случайных величин являются независимыми? Поясните.

а) X_1 и X_2

б) X_2 и X_3

в) (X_1, X_2) и X_3

г) $\frac{X_1 + X_2}{2}$ и X_3.

5. Вычислите статистику Хоттелинга для проверки гипотезы H_0: $\boldsymbol{\mu} = [7, 11]$, используя данные $\mathbf{X}^T = \begin{bmatrix} 2 & 8 & 6 & 8 \\ 12 & 9 & 9 & 10 \end{bmatrix}$. Проверьте гипотезу на 5% уровне значимости.

6. Постройте 90% доверительную область для $\boldsymbol{\mu}$ из предыдущей задачи.

7. Оцените значимость различий двух рынков сбыта бытовой техники. На первом рынке (число наблюдений — 5) средний уровень цены реализации составил 5 тыс руб., а экспертная оценка качества обслуживания — 3,4 балла; на втором рынке (число наблюдений — 7) соответственно 7 тыс руб. и 4,3 балла. Объединенная ковариационная матрица имеет вид: $\mathbf{S} = \begin{bmatrix} 9{,}3 & 0{,}26 \\ 0{,}26 & 2{,}0 \end{bmatrix}$. Принять уровень надежности — 95%.

8. Кратко поясните логическую схему построения статистического критерия для проверки однородности нормальной выборочной совокупности.

9. Каковы основные характеристики многомерной случайной величины?

10. Изобразите графически двумерную доверительную область для случайной величины (X_1, X_2), если X_1 — число часов работы в сутки, X_2 — число лет образования. Причем $\bar{X}_1 = 8{,}67, \left|\bar{X}_1 - X_{1i}\right| \le 3{,}2$ и $\bar{X}_2 = 12{,}16, \left|\bar{X}_2 - X_{2i}\right| \le 3{,}8$.

ГЛАВА 3
Корреляционный и регрессионный анализ многомерной генеральной совокупности

Бросая в воду камешки, смотри на круги, ими образуемые; иначе такое бросание будет пустою забавою.

К. Прутков. Мысли и афоризмы

Большая часть информации о поведении случайной величины скрывается в оценках средних значений многомерного признака и попарных корреляций между признаками, которые образуют корреляционную матрицу.

Корреляционный анализ многомерной генеральной совокупности решает *задачи*:

а) выбор подходящего показателя статистической связи между признаками и оценка его значения по выборке;

б) построение интервала значений для выбранного показателя (проверка его статистической значимости);

в) определение структуры связи между компонентами исследуемого многомерного признака.

Регрессионный анализ применяется в случаях, когда изучаемый процесс или явление является результатом совместного действия нескольких факторов, а у исследователя возникает потребность в оценке влияния каждого фактора в отдельности.

3.1. Корреляционный анализ количественных признаков

Рассмотрим случай трех признаков $X=(X_1, X_2, X_3)$, $(p=3)$. Будем предполагать, что поведение многомерного вектора X описывается нормальным законом распределения, т.е. плотность совместного распределения одномерных случайных величин X_1, X_2, X_3 задается в виде:

$$p(X_1,X_2,X_3)=\frac{1}{\sqrt{(2\pi)^3}}\cdot\frac{1}{\sqrt{\sigma_{X_1}^2\sigma_{X_2}^2\sigma_{X_3}^2\det(\mathbf{R})}}e^{-0{,}5z^T\mathbf{R}^{-1}z}, \quad (3.1)$$

где R — симметрическая положительно определенная матрица парных коэффициентов корреляции, а $\det(\mathbf{R})$ — определитель этой матрицы (обобщенная дисперсия случайной величины X), т.е.

$$\mathbf{R}=\begin{bmatrix} 1 & r_{X_1X_2} & r_{X_1X_3} \\ r_{X_2X_1} & 1 & r_{X_2X_3} \\ r_{X_3X_1} & r_{X_3X_2} & 1 \end{bmatrix} \text{ и}$$

$$|\mathbf{R}|=\det\mathbf{R}=1+2r_{X_1X_2}r_{X_1X_3}r_{X_2X_3}-r_{X_1X_3}^2-r_{X_2X_3}^2-r_{X_1X_2}^2>0.$$

Также в (3.1) $\mathbf{R}^{-1}$ — матрица, обратная к R, т.е.

$$R^{-1}\cdot R=I=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

В формуле (3.1) буквой $\mathbf{z}=\begin{bmatrix} \frac{X_1-\mu_{X_1}}{\sigma_{X_1}} \\ \frac{X_2-\mu_{X_2}}{\sigma_{X_2}} \\ \frac{X_3-\mu_{X_3}}{\sigma_{X_3}} \end{bmatrix}$ обозначен вектор значений нормированных случайных величин X_1, X_2, X_3.

Таким образом, имеется трехмерная нормально распределенная случайная величина, которая определяется девятью параметрами: $EX_1=\mu_{X_1}, EX_2=\mu_{X_2}, EX_3=\mu_{X_3}, VX_1=\sigma_{X_1}^2, VX_2=\sigma_{X_2}^2, VX_3=\sigma_{X_3}^2, r_{X_1X_2}, r_{X_1X_3}, r_{X_2X_3}$.

Распределения одномерных X_1, X_2, X_3, двумерных (X_1,X_2), (X_1,X_3), (X_2,X_3), условные распределения при фиксировании одной из переменных $(X_1,X_2)|X_3$; $(X_1,X_3)|X_2$, $(X_2,X_3)|X_1$ и двух $X_1|X_2X_3$, $X_2|X_1X_3$, $X_3|X_2X_1$ являются нормальными.

Для многомерной корреляционной модели важную роль играют частные и множественные коэффициенты корреляции, детерминации (квадраты соответствующих коэффициентов корреляции).

Частный коэффициент корреляции между X_1 и X_2 при фиксированном воздействии переменной X_3 может быть определен по следующей формуле:

$$r_{X_1X_2}(X_3)=\frac{-R_{12}}{\sqrt{R_{11}R_{12}}}=\frac{r_{X_1X_2}-r_{X_1X_3}r_{X_2X_3}}{\sqrt{(1-r_{X_1X_3}^2)(1-r_{X_2X_3}^2)}}, \quad (3.2)$$

где R_{ij} — алгебраическое дополнение матрицы $\mathbf{R}$ к элементу r_{ij}.

Частный коэффициент корреляции показывает тесноту линейной связи между двумя переменными случайными величинами независимо от влияния остальных случайных величин.

Обладает всеми свойствами парного коэффициента корреляции.

Если частный коэффициент корреляции (3.2) меньше парного, т.е. $r_{X_1X_2}(X_3)<r_{X_1X_2}$, то взаимодействие между X_1 и X_2 обусловлено частично (или полностью, если $r_{X_1X_2}(X_3)=0$) воздействием фиксируемых прочих переменных, т.е. — X_3. Если частный коэффициент корреляции $r_{X_1X_2}(X_3)>r_{X_1X_2}$, то фиксируемые прочие переменные ослабляют линейную связь.

Аналогично (3.2): $$r_{X_1X_3}(X_2)=\frac{r_{X_1X_3}-r_{X_1X_2}r_{X_3X_2}}{\sqrt{(1-r_{X_1X_2}^2)(1-r_{X_3X_2}^2)}}. \quad (3.3)$$

$$r_{X_2X_3}(X_1)=\frac{r_{X_2X_3}-r_{X_2X_1}r_{X_1X_3}}{\sqrt{(1-r_{X_2X_1}^2)(1-r_{X_1X_3}^2)}}. \quad (3.4)$$

Множественный коэффициент корреляции между одной переменной и другими может быть рассчитан по формуле:

$$r_{x_3}=r_{X_3}(X_1X_2)=\sqrt{1-\frac{\det\mathbf{R}}{R_{11}}}=\sqrt{\frac{r_{X_3X_1}^2+r_{X_3X_2}^2-2r_{X_1X_2}r_{X_3X_1}r_{X_3X_2}}{1-r_{X_1X_2}^2}}. \quad (3.5)$$

Если $r_{X_3}=1$, то точки (X_1, X_2, X_3) расположены в плоскости регрессии X_3 на (X_1, X_2), т.е. имеется линейная связь между пе-

ременными X_3 и двумерной переменной (X_1, X_2). Если $r_{X_3} = 0$, то линейной связи нет. Аналогично:

$$r_{X_2} = r_{X_2}(X_1X_3) = \sqrt{\frac{r^2_{X_2X_1} + r^2_{X_2X_3} - 2r_{X_2X_1}r_{X_2X_3}r_{X_1X_3}}{(1-r^2_{X_2X_3})}} \text{ и}$$

$$r_{X_1} = r_{X_1}(X_2X_3) = \sqrt{\frac{r^2_{X_1X_2} + r^2_{X_1X_3} - 2r_{X_1X_2}r_{X_1X_3}r_{X_2X_3}}{(1-r^2_{X_2X_3})}}.$$

Множественный коэффициент детерминации (квадрат соответствующего множественного коэффициента корреляции) показывает долю дисперсии, например, случайной величины X_3, обусловленную изменением случайных величин X_1 и X_2.

Проверка статистической значимости множественного коэффициента корреляции.

Нулевая гипотеза: отсутствует линейная связь между переменной x и остальными переменными, образующими многомерный признак, H_0: $r_x=0$, H_1: $r_x \neq 0$.

Рассчитываем статистику

$$F = \frac{r_x^2/2}{(1-r_x^2)/(n-3)}.$$

Если F больше значения распределения Фишера $F_\varepsilon(2, n-3)$, то гипотеза H_0 отвергается, следовательно, линейная связь есть и она статистически значима на $\varepsilon 100\%$ уровне значимости.

Интервальная оценка для частного коэффициента корреляции. Для получения интервальной оценки частного коэффициента корреляции используется z статистика Фишера:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r} \text{ и } \Delta z = u_\gamma \frac{1}{\sqrt{n-4}},$$

где u_γ является решением уравнения $\Phi(u_\gamma) = \gamma$ и находится по таблице интегральной функции Лапласа. Тогда $(z_r \pm \Delta z)$ — интервал для $E(z)$ и доверительный интервал для частного коэффициента корреляции получают по таблице обратного преобразования Фишера.

Интервальная оценка для множественного коэффициента корреляции находится по графикам и таблицам с помощью z-преобразования Фишера.

3.2. Ранговая корреляция

При изучении неколичественных признаков или количественных с непрерывными и неизвестными законами распределения классический подход корреляционного анализа оказывается неэффективен, и в этих случаях применяют методы непараметрической статистики и, в частности, метод ранговой корреляции.

Ранговая корреляция предназначена для изучения статистической связи между различными упорядочиваниями (ранжировками) объектов по степени проявления в них того или иного свойства.

Пусть $X^{(j)} = (x_1^{(j)}, x_2^{(j)}, ..., x_n^{(j)})$ некоторая ранжировка из n объектов по j-му свойству. Компонента $x_i^{(j)}$, i=1,...,n этой ранжировки определяет порядковое место (ранг), которое присвоено i-му объекту в общем ряду n анализируемых объектов, упорядоченных по убыванию j-го свойства. Другая ранжировка $X^{(k)} = (x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)})$ может интерпретироваться либо как упорядочивание объектов по другому k-му свойству, либо как ранжировка по тому же свойству, но полученная, например, другим экспертом.

Ранговый коэффициент корреляции Спирмена.

Ранговый коэффициент корреляции Спирмена измеряет степень согласованности двух различных ранжировок $X^{(j)}$ и $X^{(k)}$ одного и того же множества из n объектов и рассчитывается по формуле

$$r_c = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}(x_i^{(j)} - x_i^{(k)})^2.$$

Причем $-1 \le r_c \le 1$. Соответственно -1 означает, что ранжировки противоположны, 1 — они совпадают, r_c=0 — между ранжировками связь отсутствует.

Связанные ранги возникают в случае дележки мест в ранжировке. Объектам, которые делят места, приписывается ранг, равный среднему арифметическому соответствующих мест. Например: объекты делят места с 3 по 5, тогда $\frac{3+4+5}{3} = 4$ и объектам на 3, 4 и 5 местах приписывается ранг 4.

Для связанных рангов формула коэффициента корреляции Спирмена усложняется: $r_c = \dfrac{\frac{1}{6}(n^3-n) - \sum_{i=1}^{n}(x_i^{(j)} - x_i^{(k)})^2 - T^{(j)} - T^{(k)}}{\left[\frac{1}{6}(n^3-n) - 2T^{(j)}\right]^{\frac{1}{2}}\left[\frac{1}{6}(n^3-n) - 2T^{(k)}\right]^{\frac{1}{2}}}$,

где $T^{(l)} = \frac{1}{12}\sum_{t=1}^{m}(n_t^{(l)})^3 - n_t^{(l)}$, здесь m — число групп, для которых имеются связанные ранги, n_t — число рангов, входящих в t-ю группу.

Проверка гипотезы H_0: $r_c=0$. Если $n>10$, то статистика $\gamma_n = \dfrac{r_c\sqrt{n-2}}{\sqrt{1-r_c^2}}$ распределена по закону Стьюдента с $(n-2)$ степенями свободы. При $\gamma_n > t_{\alpha/2}(n-2)$ гипотеза H_0 отклоняется, иначе — не отклоняется. Если $n\le 10$, то рассчитывают $r_c^{\max} = \dfrac{2S_c}{(n^3-n)/3}$, где S_c извлекается из таблицы для уровня значимости $\alpha/2$ и если $|r_c| > r_c^{\max}$, то $\mathbf{H}_0$ отвергается.

Ранговый коэффициент корреляции Кендалла. Расчетная формула имеет вид: $r_k = \dfrac{S}{\frac{1}{2}n(n-1)} = \dfrac{P-Q}{P+Q}, S = P-Q$.

Ранжируем все элементы по признаку $x^{(1)}$, по ряду другого признака $x^{(2)}$ подсчитываем для каждого ранга число последующих рангов, превышающих данный (их сумму обозначим P) и число последующих рангов ниже данного (их сумму обозначим Q). Значения коэффициента $-1 \le r_k \le 1$.

На практике при $n \ge 10$, $r_c \approx \dfrac{3}{2}r_k$. Для связанных рангов формула коэффициента корреляции Кендалла имеет вид

$$r_k^* = \frac{r_k - \dfrac{2(u^{(1)}+u^{(2)})}{n(n-1)}}{\sqrt{\left(\left(1-\dfrac{2u^{(1)}}{n(n-1)}\right)\left(1-\dfrac{2u^{(2)}}{n(n-1)}\right)\right)}}, u^{(l)} = \frac{1}{2}\sum_{t=1}^{m^{(l)}} n_t^{(l)}(n_t^{(l)}-1), l = 1,2.$$

Коэффициент конкордации (согласованности) Кендалла. Измеряет степень тесноты, статистической связи между m различными ранжировками:

$$W = \frac{12}{m^2(n^3-n)}\sum_{i=1}^{n}\left(\sum_{l=1}^{m} x_i^{(l)} - \frac{m(n+1)}{2}\right)^2.$$

Причем $0 \le W \le 1$, если $W=1$ — ранжировки совпадают, $W=0$ — связь между ранжировками отсутствует.

Для связанных рангов $W = \dfrac{\sum_{i=1}^{n}\left(\sum_{l=1}^{m}(x_i^{(l)} - \frac{m(n+1)}{2}\right)^2}{\frac{1}{12}m^2(n^3-n) - m\sum_{l=1}^{m}T^{(l)}}$.

Проверка гипотезы H_0: $W=0$, при $n>7$ рассчитывают статистику $\gamma_n = m(n-1)W$ и при $\gamma_n > \chi_\alpha^2(n-1)$ гипотеза отклоняется, иначе не отклоняется.

3.3. Корреляция категоризованных переменных

Признак называется категоризованным, если его возможные «значения» описываются конечным числом состояний (категорий, градаций). В этом случае для двух категоризованных признаков исходные данные представляются в виде *таблиц сопряженности*.

Градации признака $x^{(1)}$	Градации признака $x^{(2)}$						n_i
	1	2	...	j	...	m_2	
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1m_2}	$n_{1\cdot}$
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2m_2}	$n_{2\cdot}$
⋮	...	...		...	...	...	...
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{im_2}	$n_{i\cdot}$
⋮		...		...			
m_1	$n_{m1}1$	$n_{m1}2$	...	n_{m_1j}	...	$n_{m_1m_2}$	$n_{m_1\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot j}$	...	$n_{\cdot m_2}$	n

В таблице n_{ij} — число объектов (из общего числа n обследованных), у которых «значение» признака $x^{(1)}$ на i-м уровне градации, а «значение» признака $x^{(2)}$ на j-м уровне. Также $n_{i.} = \sum_{j=1}^{m_2} n_{ij}$ — число объектов, значение признака $x^{(1)}$ у которых на i-й градации (i=1, 2, ..., m_1); $n_{.j} = \sum_{i=1}^{m_1} n_{ij}$ — число объектов, значение признака $x^{(2)}$ у которых на j-й градации (j=1, 2, ..., m_2).

Основные измерители степени тесноты статистической связи между категоризованными переменными.

1. Коэффициент квадратической сопряженности рассчитывается по формуле:

$$\chi^2 = n\left(\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right). \qquad (3.6)$$

Причем $0 \le \chi^2 < +\infty$, чем больше χ^2 отличается от 0, тем сильнее связь.

Проверка гипотезы $H_0 : \chi^2 = 0$. Если $\chi^2 \ge \chi^2_\alpha((m_1-1)(m_2-1))$, то гипотеза отклоняется, иначе не отклоняется.

Вместо (3.6) часто используют коэффициент Крамера, значения которого принадлежат отрезку от 0 до 1:

$$c = \left(\frac{\chi^2}{n\min(m_1-1, m_2-1)}\right)^{1/2}.$$

2. Информационная мера связи (другое название — отношение правдоподобия) $Y^2 = 2\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} n_{ij}\ln\left(\frac{n_{ij}}{n_{i.}n_{.j}/n}\right)$.

Гипотеза $H_0 : Y^2 = 0$ также проверяется на основе статистики $\chi^2_\alpha\left((m_1-1)(m_2-1)\right)$.

Пирсон предложил для измерения связи показатель сопряженности ($0 \le C \le 1$) $C = \sqrt{\frac{\chi^2}{n+\chi^2}}$.

Его недостаток в том, что он не достигает единицы и при полной связи признаков, а лишь стремится к единице.

Также известен показатель связи Чупрова: $T = \sqrt{\frac{\chi^2}{n\sqrt{(m_1-1)(m_2-1)}}}$, который строже оценивает тесноту связи, чем показатель Пирсона.

На практике часто по таблице сопряженности необходимо проверить взаимную независимость двух переменных. Эту задачу обычно решают с помощью критерия Пирсона

$\chi^2 = \sum_i\sum_j \frac{\left(n_{ij} - n_{ij}^e\right)^2}{n_{ij}^e}$, где n_{ij}^e — ожидаемые частоты.

При справедливости нуль-гипотезы о независимости и при достаточно большом объеме выборки статистика χ^2 распределена в соответствии с хи-квадрат распределением при числе степеней свободы $(m_1-1)(m_2-1)$.

Другой формой критерия хи-квадрат является тест Мантеля-Хензеля, который определяется как произведение парного коэффициента корреляции на количество наблюдений, уменьшенное на единицу: $\chi^2_{MH} = r^2_{x^{(1)}x^{(2)}}(n-1)$.

В тесте число степеней свободы равно 1.

Рассмотрим частный случай таблицы сопряженности 2×2, когда $m_1 = m_2 = 2$.

$x^{(1)}$	$x^{(2)}$		Итого
	1	2	
1	n_{11}	n_{12}	A
2	n_{21}	n_{22}	B
Итого	a	b	A+B=a+b

В таблице $a = n_{11}+n_{21}$, $A = n_{11}+n_{12}$, $b = n_{12}+n_{22}$, $B = n_{21}+n_{22}$. Тогда может быть рассчитан коэффициент ассоциации Пирсона

$$(0 \le Q \le 1)\ Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{abAB}}.$$

Свойства у коэффициента ассоциации такие же, как и у коэффициента корреляции.

Коэффициент контингенции Юла-Кендалла

$$Q_k = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}.$$

Значение коэффициента контингенции выше, чем коэффициента ассоциации. Недостатком коэффициента Юла-Кендалла является то, что он завышает тесноту связи.

Только для таблиц 2×2 часто рассчитывается также мера связи, предложенная Фишером $\varphi = \sqrt{\frac{\chi^2}{n}}$, поскольку в других случаях его значение может превысить 1.

3.4. Регрессионный анализ

Всякий раз, когда изучаемый процесс или явление является результатом совместного действия нескольких факторов, у исследователя возникает потребность в оценке влияния каждого фактора в отдельности. Один из стандартных методов*, позволяющий успешно решить эту задачу, суть *множественная регрессия*.

Пусть мы располагаем выборочными наблюдениями над переменными Y_i и X_{ji}, j=1,..., p, i=1,2,...,n, где n — количество наблюдений:

1	2	...	i	...	n
Y_1	Y_2	...	Y_i	...	Y_n
X_{11}	X_{12}	...	X_{1i}	...	X_{1n}
...	...	...	...	...	...
X_{p1}	X_{p2}	...	X_{pi}	...	X_{pn}

Предположим, что существует линейное соотношение между результирующей переменной Y и p объясняющими переменными $X_1, X_3, ..., X_p$. Тогда с учетом случайной ошибки u_i запишем уравнение:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + u_i, \quad i = 1,2,...,n \qquad (3.7)$$

* Другой возможный путь решения - это известная схема управляемого эксперимента — см., например: *Джонсон Н., Лион Ф.* Статистика и планирование эксперимента в технике и науке. В 2-х т. — М.: Мир, 1980.

В (3.7) неизвестны коэффициенты β_j, j=0,2,...,p и параметры распределения u_i. Задача состоит в оценивании этих неизвестных величин. Модель (3.7) называется *классической линейной моделью множественной регрессии* (КЛММР). Заметим, что часто имеют в виду, что переменная X_0 при β_0 равна единице для всех наблюдений i=1,2,...,n.

Относительно переменных модели в уравнении (3.7) примем следующие основные гипотезы:

$$E(u_i) = 0; \qquad (3.8)$$

$$E(u_i u_j) = \begin{cases} \sigma^2 & \text{при } i = j, \\ 0 & \text{при } i \neq j, \end{cases} \qquad (3.9)$$

$X_1, X_3, ..., X_p$ — неслучайные переменные. (3.10)

Не должно существовать строгой линейной зависимости между переменными $X_1, X_3, ..., X_p$. (3.11)

Первая гипотеза (3.8) означает, что переменные u_i имеют нулевую среднюю. Суть гипотезы (3.9) в том, что все случайные ошибки u_i имеют постоянную дисперсию, т.е. выполняется условие *гомоскедастичности* дисперсии.

Согласно (3.10) в повторяющихся выборочных наблюдениях источником возмущений Y являются случайные колебания u_i, а значит, свойства оценок и критериев обусловлены объясняющими переменными $X_1, X_3, ..., X_p$.

Последняя гипотеза (3.11) означает, в частности, что не существует *линейной зависимости* между объясняющими переменными, включая переменную X_0, которая всегда равна 1.

Обозначим:

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{bmatrix}; \quad \underset{n\times p}{\mathbf{X}} = \begin{bmatrix} X_{11} & ... & X_{1p} \\ X_{21} & ... & X_{2p} \\ & ... & \\ X_{n1} & ... & X_{np} \end{bmatrix}; \quad \underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_2 \\ ... \\ \beta_p \end{bmatrix}; \quad \underset{n\times 1}{\mathbf{u}} = \begin{bmatrix} u_1 \\ u_2 \\ ... \\ u_n \end{bmatrix}.$$

Тогда модель (3.7)-(3.11) запишется в матричном виде: $\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$, при ограничениях:

$$E(\mathbf{u}) = \mathbf{0}; \qquad (3.8a)$$

$$E(\mathbf{uu}^T) = \mathbf{\Sigma}^2\mathbf{I}; \qquad (3.9a)$$

$\mathbf{X}$ — детерминированная матрица; (3.10а)

$\text{rang}(\mathbf{X})=p<n$. (3.11а)

Применяя к (3.7) с учетом (3.8)-(3.11) метод наименьших квадратов (МНК), получаем из необходимых условий минимизации функционала: $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - ... - \hat{\beta}_p X_{pi})^2$, т.е. обращения в нуль частных производных по каждому из параметров:

$$\frac{\partial}{\partial \hat{\beta}_0}\left(\sum_{i=1}^{n} e_i^2\right) = -2\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - ... - \hat{\beta}_p X_{pi}) = 0;$$

$$\frac{\partial}{\partial \hat{\beta}_1}\left(\sum_{i=1}^{n} e_i^2\right) = -2\sum_i X_{1i}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - ... - \hat{\beta}_p X_{pi}) = 0;$$

...

$$\frac{\partial}{\partial \hat{\beta}_k}\left(\sum_{i=1}^{n} e_i^2\right) = -2\sum_i X_{pi}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - ... - \hat{\beta}_p X_{pi}) = 0.$$

Упростив последние равенства, получим стандартную форму *нормальных уравнений*, решение которых дает искомые оценки параметров:

$$\begin{cases} \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i} + ... + \hat{\beta}_p \sum_{i=1}^{n} X_{pi}; \\ \sum_{i=1}^{n} Y_i X_{1i} = \hat{\beta}_0 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_{1i} X_{2i} + ... + \hat{\beta}_p \sum_{i=1}^{n} X_{1i} X_{pi}; \\ ... \\ \sum_{i=1}^{n} Y_i X_{pi} = \hat{\beta}_0 \sum_{i=1}^{n} X_{pi} + \hat{\beta}_1 \sum_{i=1}^{n} X_{pi} X_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{pi} X_{2i} + ... + \hat{\beta}_p \sum_{i=1}^{n} X_{pi}^2. \end{cases} \quad (3.12)$$

Сложность решения системы линейных уравнений (3.12) с $(k+1)$ неизвестными увеличивается быстрее, чем растет p. В зависимости от количества уравнений система может быть решена методом исключения Гаусса или методом Крамера или другим численным методом решения системы линейных алгебраических уравнений.

Поскольку для большинства практических задач изучаются несколько альтернативных спецификаций модели (3.7), то

широкое применение ЭВМ, а также специальных статистических пакетов позволяет значительно упростить процедуру оценивания.

В результате решения системы (3.12) получим оценки коэффициентов $\hat{\beta}_j$, $j=0,2,...,p$.

В матричной нотации сумма квадратов отклонений будет равна: $\sum e_i^2 = \mathbf{e}^T\mathbf{e} = \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)^T\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right) = \mathbf{y}^T\mathbf{y} - 2\hat{\beta}^T\mathbf{X}^T\mathbf{y} + \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta}$.

Для нахождения $\hat{\beta}$ последнее равенство необходимо продифференцировать по $\hat{\beta}$, т.е. $\frac{\partial}{\partial \hat{\beta}}\left(\mathbf{e}^T\mathbf{e}\right) = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\beta}$, откуда, приравнивая к нулю, получим

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. \quad (3.13)$$

Формула (3.13) дает выражение для оценок КЛММР методом наименьших квадратов.

Возможна и другая запись уравнения (3.7), в так называемом *стандартизованном* масштабе:

$$t_Y = b_1 t_{X_1} + b_2 t_{X_2} + ... + b_p t_{X_p} + u, \quad (3.14)$$

где $t_Y, t_{X_1}, ..., t_{X_p}$ — стандартизованные переменные: $t_Y = \frac{Y - \bar{Y}}{\sigma_Y}$, $t_{X_j} = \frac{X_j - \bar{X}_j}{\sigma_{X_j}}$, $j=1,2,...,p$, для которых среднее значение равно нулю: $\bar{t}_Y = \bar{t}_{X_j} = 0$, $j=1,2,...,p$, а среднее квадратическое отклонение равно единице: $\sigma_{t_Y} = \sigma_{t_{X_j}} = 1$, $j=1,2,...,p$, b_j, $j=1,2,...,p$ — *стандартизованные коэффициенты регрессии.*

Нетрудно установить зависимость между коэффициентами «чистой» регрессии β_j и стандартизованными коэффициентами регрессии b_j, $j=1,2,...,p$, а именно:

$$b_j = \beta_j \frac{\sigma_{X_j}}{\sigma_Y}, \; j=1,2,...,p, \quad (3.15)$$

причем $\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - ... - \beta_p \bar{X}_p$.

Соотношение (3.15) позволяет переходить от уравнения вида (3.14) к уравнению вида (3.7).

Стандартизованные коэффициенты регрессии показывают, на сколько «сигм» изменится в среднем результат (Y), если

соответствующий фактор X_j изменится на одну «сигму» при неизменном среднем уровне других факторов.

В силу того, что все переменные центрированы и нормированы, коэффициенты b_j, j=1,2,...,p, сравнимы между собой (в этом их отличие от β_j). Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат, что позволяет произвести отсев факторов — исключить из модели факторы с наименьшими значениями b_j.

Оценки МНК, полученные по формуле (3.13) в условиях (3.8а)-(3.11а), являются наиболее эффективными (в смысле наименьшей дисперсии) оценками в классе линейных несмещенных оценок (теорема Гаусса-Маркова).

Как было уже указано раньше, достоинством метода множественной регрессии является возможность выделения влияния каждого из факторов X_j в условиях, когда воздействие многих переменных на результат эксперимента не удается контролировать. Степень раздельного влияния каждого из факторов характеризуется оценками $\hat{\beta}_j$, j=1,2,...,p.

Можно получить также полезную формулу для суммы квадратов остатков КЛММР: $\mathbf{e}^T\mathbf{e} = \mathbf{y}^T\mathbf{y} - \hat{\beta}^T\mathbf{X}^T\mathbf{y}$.

Опуская вывод выражения для дисперсий коэффициентов, образующих вектор $\hat{\beta}$, приведем его результат:

$$V(\hat{\beta}) = \sigma\left(\mathbf{X}^T\mathbf{X}\right)^{-1}. \qquad (3.16)$$

Матрицу в (3.16) называют *ковариационной матрицей*, в которой на главной диагонали находятся дисперсии оценок $\hat{\beta}_j$, а элементы с индексами *ij* содержат ковариации пар оценок $\hat{\beta}_i$ и $\hat{\beta}_j$.

Для дисперсии возмущающего воздействия может быть получена несмещенная оценка: $\hat{\sigma}^2 = \dfrac{\mathbf{e}^T\mathbf{e}}{n-p-1}$. (3.17)

Заметим, что множественный коэффициент корреляции может быть вычислен с использованием стандартизованных коэффициентов регрессии b_{X_i} и парных коэффициентов корреляции r_{YX_i} $R_{YX_1X_2\ldots X_k} = \sqrt{\sum b_{X_i} \cdot r_{YX_i}}$.

Также коэффициент детерминации R^2 свидетельствует о качестве регрессионной модели и отражает долю общей вариации результирующего признака Y, объясненную изменением функции регрессии, и может быть найден по формуле

$$R^2 = 1 - \frac{\hat{\sigma}^2}{\sigma_Y^2} = 1 - \frac{\mathbf{e}^T\mathbf{e}}{\mathbf{y}^T\mathbf{y}}. \qquad (3.18)$$

Однако использование R^2 в случае множественной регрессии является *не вполне корректным*, так как коэффициент детерминации возрастает при добавлении регрессоров в модель. Это происходит потому, что остаточная дисперсия уменьшается при введении дополнительных переменных. И если число факторов приблизится к числу наблюдений, то остаточная дисперсия будет равна нулю, и коэффициент множественной корреляции, а значит и коэффициент детерминации, приблизятся к единице, хотя в действительности связь между факторами и результатом и объясняющая способность уравнения регрессии могут быть значительно ниже.

Для того чтобы получить адекватную оценку того, насколько хорошо вариация результирующего признака объясняется вариацией нескольких факторных признаков, применяют *скорректированный коэффициент детерминации*

$$R^2_{\text{скорр}} = 1 - (1 - R^2) \cdot \frac{n-1}{n-p-1}. \qquad (3.19)$$

Скорректированный коэффициент детерминации всегда меньше R^2. Кроме того, в отличие от R^2, который всегда положителен, $R^2_{\text{скорр}}$ может принимать и отрицательное значение.

Проверка гипотез в модели множественной регрессии.

Для проверки значимости коэффициентов КЛММР необходимо к предположениям (3.8)-(3.11) добавить предположение о нормальном распределении остатков u_i в (3.7). Это позволяет воспользоваться t-распределением для проверки гипотез относительно каждого из регрессионных коэффициентов $\hat{\beta}_j \in N\left(\beta_j, \sigma^2_{\hat{\beta}_j}\right)$.

Для проверки гипотезы H_0: $\beta_j = \beta_{j0}$ используется статистика:

$$t = \frac{\hat{\beta}_j - \beta_{j0}}{\mu_{\hat{\beta}j}}, \qquad (3.20)$$

где $\mu_{\hat{\beta}_j}$ — стандартная ошибка коэффициента $\hat{\beta}_j$, которая может быть вычислена по формуле с учетом (3.16) и результата (3.17): $\mu_{\hat{\beta}_j} = \sqrt{q_{jj}\hat{\sigma}^2} = \sqrt{q_{jj}}\sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2}$, здесь q_{jj} — j-й диагональный элемент матрицы $(\mathbf{X}^T\mathbf{X})^{-1}$.

Поскольку t в (3.20) распределено по Стьюденту, то, сравнивая с критическим значение $t_{\alpha/2}(n-p-1)$* для заданного уровня значимости α, принимаем нулевую гипотезу, если $|t| < t_{\alpha/2}(n-p-1)$, и отклоняем в противном случае.

Ясно, что с помощью (3.20) можно проверить гипотезу о значимости отдельного коэффициента регрессии H_0: $\beta_j = 0$, здесь $\beta_{j0} = 0$.

Также можно получить $100(1-\varepsilon)\%$ доверительный интервал для коэффициента множественной регрессии β_j:

$$\beta_j = \hat{\beta}_j \pm t_{\alpha/2}(n-p-1)\mu_{\hat{\beta}_j}.$$

Для осуществления проверки гипотез относительно нескольких или всех β_j используют дисперсионный анализ.

Пусть необходимо проверить гипотезу H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$.

Сумма квадратов отклонений от среднего в выборке равна сумме квадратов отклонений значений $\hat{Y}$, полученных по уравнению регрессии, от выборочного среднего $\bar{Y}$ плюс сумма квадратов отклонений Y от линии регрессии $\hat{Y}$, т.е. объясненная часть (RSS — regression sum of squares) суть $\sum y_i^2 - \sum e_i^2 = \hat{\beta}^T\mathbf{X}^T\mathbf{y} - \frac{1}{n}\left(\sum Y\right)^2$, а необъясненный остаток (ESS — error sum of squares) $\sum e_i^2 = \mathbf{e}^T\mathbf{e} = \mathbf{y}^T\mathbf{y} - \hat{\beta}^T\mathbf{X}^T\mathbf{y}$. Если переменные выражены в виде отклонений от своих средних значений, то $\sum y_i^2 - \sum e_i^2 = \hat{\beta}^T\mathbf{X}^T\mathbf{y}$.

С учетом (3.18) получим таблицу дисперсионного анализа (табл. 3.1).

Тогда проверка гипотезы H_0 (линейная связь между $X_1, X_2, \ldots, X_p$ и Y отсутствует) может быть осуществлена с помощью F-критерия Фишера. Для проверки гипотезы воспользуемся соотношением:

$$F = \frac{D_1}{D_2} = \frac{\hat{\beta}^T\mathbf{X}^T\mathbf{y}/p}{\mathbf{e}^T\mathbf{e}/(n-p-1)} = \frac{R^2}{p} : \frac{1-R^2}{n-p-1}, \qquad (3.21)$$

которое удовлетворяет F-распределению Фишера с $(p, n-p-1)$ степенями свободы. Критические значения этой статистики F_α для уровня значимости α затабулированы.

Таблица 3.1

Таблица дисперсионного анализа

Источник вариации	Сумма квадратов отклонений	Число степеней свободы	Средняя сумма квадратов
$X_1, X_2, \ldots, X_p$	$Q_1 = \sum\left(\hat{Y} - \bar{Y}\right)^2 = \hat{\beta}^T\mathbf{X}^T\mathbf{y} = \mathbf{y}^T\mathbf{y} \cdot R^2$	p	$D_1 = \frac{Q_1}{p}$
Остаток	$Q_2 = \sum\left(Y - \hat{Y}\right)^2 = \mathbf{e}^T\mathbf{e} = \mathbf{y}^T\mathbf{y}(1-R^2)$	$n-p-1$	$D_2 = \frac{Q_2}{n-p-1}$
Общая вариация	$\sum(Y-\bar{Y})^2 = Q_1 + Q_2 = \mathbf{y}^T\mathbf{y}$	$n-1$	

Если $F > F_\alpha$, то гипотеза об отсутствии связи между переменными $X_1, X_2, \ldots, X_p$ и Y отклоняется, в противном случае гипотеза H_0 принимается и уравнение регрессии не значимо.

Линейное ограничение общего вида H_0: $\mathbf{H}\beta = \mathbf{r}$. Пусть $\mathbf{H}$ — матрица $q \times p$, β — вектор коэффициентов $p \times 1$, $\mathbf{r}$ — вектор $q \times 1$. Предполагаем, что число ограничений на коэффициенты не превышает количество самих коэффициентов $q \le p$, и все они линейно независимы, т.е. $\text{rang}(\mathbf{H}) = q$.

Пример. Пусть для модели $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$ надо проверить гипотезы $\begin{cases} \beta_1 = \beta_3, \\ \beta_2 = 2. \end{cases}$

Тогда в матричном виде ограничения запишутся так:

$$\mathbf{H}\beta = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \mathbf{r}.$$

* Здесь и далее p — число регрессоров без константы, поэтому число степеней свободы надо уменьшить на 1, чтобы учесть константу в уравнении регрессии.

Статистика

$$F=\frac{\left(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{r}\right)^T\left(\mathbf{H}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{H}^T\right)^{-1}\left(\mathbf{H}\hat{\boldsymbol{\beta}}-\mathbf{r}\right)/q}{\mathbf{e}^T\mathbf{e}/(n-p-1)} \qquad (3.22)$$

имеет распределение Фишера с q и $n-p-1$ степенями свободы. Так, что при выполнении неравенства $F_\alpha(q, n-p-1)<F$ гипотеза H_0 отклоняется, иначе — принимается.

Если, в частности, $\mathbf{H}=\boldsymbol{I}$, то формула (3.22) упрощается:

$$F=\frac{\left(\hat{\boldsymbol{\beta}}-\mathbf{r}\right)^T\left(\mathbf{X}^T\mathbf{X}\right)\left(\hat{\boldsymbol{\beta}}-\mathbf{r}\right)/(p+1)}{\mathbf{e}^T\mathbf{e}/(n-p-1)}\sim F(p+1,n-p-1).$$

Часто необходимо протестировать гипотезу о том, что некоторое подмножество коэффициентов одновременно статистически незначимо. Т.е. имеем частный случай общей линейной гипотезы вида: $H_0:\beta_{p-q+1}=\beta_{p-q+2}=\ldots=\beta_p=0$. Значит, q коэффициентов равны нулю (остальных коэффициентов, не задействованных в гипотезе, $p-q$ штук).

Построим «короткую» регрессию зависимой переменной на $p-q$ штук переменных-факторов, не входящих в ограничение, и получим остатки этой регрессии $\mathbf{e}^*$. Также построим «длинную» регрессию зависимой переменной на все k переменных-факторов и получим остатки этой регрессии $\mathbf{e}$.

Тогда статистика:

$$F=\frac{(\mathbf{e}^{*T}\mathbf{e}^*-\mathbf{e}^T\mathbf{e})/q}{\mathbf{e}^T\mathbf{e}/(n-p-1)}=\frac{(ESS_R-ESS_{UR})/q}{ESS_{UR}/(n-p-1)}, \qquad (3.23)$$

где ESS_R — сумма квадратов остатков «короткой» (restricted) регрессии, а ESS_{UR} — сумма квадратов остатков «длинной» (unrestricted) регрессии имеет распределение Фишера с q и $n-p-1$ степенями свободы. Гипотеза принимается, если $F_\alpha(q, n-p-1)>F$.

Отметим, что статистику (3.23) можно выразить через коэффициенты детерминации «короткой» и «длинной» регрессий: $F=\dfrac{(R^2_{UR}-R^2_R)/q}{(1-R^2_{UR})/(n-p-1)}$.

Наконец рассмотрим еще один частный случай, когда проверяется гипотеза вида: H_0: $\mathbf{c}^T\boldsymbol{\beta}=\theta$, где $\mathbf{c}$ — вектор размерности $p\times1$. Этот случай получается из (3.22), когда $\mathbf{H}=\mathbf{c}^T$.

Статистика

$$t=\frac{\mathbf{c}^T\hat{\boldsymbol{\beta}}-\theta}{\sqrt{\hat{\sigma}^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}} \qquad (3.24)$$

имеет t распределение Стьюдента с $n-p-1$ степенями свободы. Если $t_\alpha(n-p-1)<t$, то гипотеза отклоняется.

Пример 3.1. Файл данных msm.sta содержит наблюдения по индивидам (приложение 1). Выберем переменные lnwage — логарифм заработной платы, expr — опыт работы на рынке труда, edu — число лет образования. Рассчитать парные и множественные коэффициенты корреляции.

Решение.

Объем выборки составил 3869 наблюдений после удаления пропущеных значений. Воспользуемся для расчета парных коэффициентов корреляции пакетом Statistica. Последовательно выбирая Statistics/Basic Statistics/Correlation matrices и указывая в качестве переменных все три имеющиеся, получим корреляционную матрицу из парных коэффициентов корреляции: $\mathbf{R}=\begin{bmatrix} \overset{lnwage}{1} & \overset{expr}{-0,20} & \overset{edu}{0,31} \\ -0,20 & 1 & -0,43 \\ 0,31 & 0,43 & 1 \end{bmatrix}$. Проверка значимости коэффициентов корреляции показала их отличие от нуля на 95% уровне надежности.

Множественный коэффициент корреляции между заработной платой и остальными переменными вычисляется как

$$r_{\ln wage}=r_{\ln wage}(edu,\exp r)=\sqrt{1-\frac{\det\mathbf{R}}{R_{11}}}.$$

Имеем

$$\det\mathbf{R}=1+2(-0,2)\cdot0,31\cdot(-0,43)-(-0,2)^2-0,31^2-(-0,43)^2=0,73.$$

Далее $R_{11}=(-1)^{1+1}\begin{vmatrix} 1 & -0,43 \\ -0,43 & 1 \end{vmatrix}=0,82.$

Тогда $r_{\ln wage} = \sqrt{1-\frac{0,73}{0,82}} = 0,32.$

Аналогично $r_{edu} = r_{edu}(\ln wage, \exp r) = \sqrt{1-\frac{\det \mathbf{R}}{R_{33}}} = \sqrt{1-\frac{0,73}{0,96}}$

$= 0,49$ и $r_{\exp r} = r_{\exp r}(\ln wage, \mathrm{e}du) = \sqrt{1-\frac{\det \mathbf{R}}{R_{22}}} = \sqrt{1-\frac{0,73}{0,90}} = 0,44.$

Частный коэффициент корреляции $r_{\ln wage, \exp r}(edu) = \frac{-R_{12}}{\sqrt{R_{11}R_{22}}}.$

Поскольку $R_{12} = (-1)^{1+2}\begin{vmatrix} -0,20 & -0,43 \\ -0,31 & 1 \end{vmatrix} = 0,07$, то

$r_{\ln wage, \exp r}(edu) = \frac{-0,07}{\sqrt{0,82 \cdot 0,90}} = -0,08.$ Таким образом, частный коэффициент корреляции между заработной платой и опытом работы оказывается меньше по модулю, чем парный коэффициент корреляции по модулю, равный 0,20, т.е. переменная числа лет образования усиливает линейную связь между lnwage и expr. Аналогично $r_{\ln wage, edu}(\exp r) = \frac{-R_{13}}{\sqrt{R_{11}R_{33}}} = \frac{0,22}{\sqrt{0,82 \cdot 0,96}} = 0,25$ и

$r_{\exp r, edu}(\ln wage) = \frac{-R_{23}}{\sqrt{R_{22}R_{33}}} = -0,39.$

Проверим, например, гипотезу о том, что H_0: $r_{\ln wage} = 0$, H_1: $r_{\ln wage} \neq 0$. Рассчитываем статистику $F = \frac{0,32^2/2}{(1-0,32^2)/(3869-3)} = 220,5,$ что больше $F_{0,05}(2, 3863) = 2,99$ и гипотеза H_0 отвергается, следовательно, линейная связь между логарифмом зарплаты и двумя другими переменными — опытом работы и числом лет образования — существует и статистически значима на 5 % уровне.

Пример 3.2. При исследовании рынка потребители распределяли товар в порядке возрастания (или убывания) своих потребительских предпочтений. Предположим, мы имеем 5 продуктов, которые ранжированы по порядку предпочтений от 1 до 5 в соответствии с двумя характеристиками A и B.

Характеристики для ранжирования	Продукты V W X Y Z
A	2 5 1 3 4
B	1 3 2 4 5

Определить показатели взаимосвязи между ранговыми оценками.

Решение.

Коэффициент ранговой корреляции Спирмена для пяти пар рангов $n = 5$ вычисляется как

$\sum_{i=1}^{5}\left(x_i^{(j)} - x_i^{(k)}\right)^2 = (2-1)^2+(5-3)^2+(1-2)^2+(3-4)^2+(4-5)^2 =$

$= 1+4+1+1+1 = 8$ и $r_c = 1-\frac{6 \cdot 8}{5(25-1)} = 0,6.$

То есть мы нашли умеренно сильную линейную связь.

Проверка значимости коэффициента Спирмена:

$t = 0,6\sqrt{\frac{5-1}{1-0,6^2}} = 1,5$ и так как $t_{0,05}(3) = 3,18$, то коэффициент не значим.

Рассчитаем коэффициент Кендалла. Все единицы ранжируются по признаку A; по ряду признака B подсчитывается для каждого ранга число последующих рангов, превышающих данный (их сумму обозначим P), и число последующих рангов ниже данного (их сумму обозначим Q). Тогда $S = P - Q$.

Преобразуем ряд

Характеристики для ранжирования	Продукты X V Y Z W
A	1 2 3 4 5
B	2 1 4 5 3

Тогда $S = (3-1)+(3-0)+(1-1)+(0-1) = 4$ и $r_k = \frac{4}{\frac{1}{2}5(5-1)} = 0,4.$

Если полученное значение умножить на 1,5, то получим 0,6 — значение коэффициента Спирмена.

Коэффициент конкордации Кендалла для случая $m=2$, $n=5$ получается как

$$W=\frac{12}{2^2(5^3-5)}[(3-6)^2+(8-6)^2+(3-6)^2+(7-6)^2+(9-6)^2]=0{,}8.$$

Пример 3.3. По результатам ответов на вопросы об уровне профессионализма и социальном слое, к которому себя относит индивид, содержащихся в файле данных msm.sta (приложение 1) сформирована таблица сопряженности. Необходимо выполнить анализ степени тесноты статистической связи между категоризованными переменными.

Уровень Вашего профессионализма	К какому слою Вы себя относите?							Всего
	Элита	Верхний слой	Выше среднего	Средний слой	Ниже среднего	Низший слой	Социальное дно	
Низкий	0	0	5	62	61	56	8	192
2	0	0	3	62	113	81	6	265
3	0	2	11	162	182	64	2	423
4	0	0	13	302	224	62	3	604
5	0	4	44	584	299	73	5	1009
6	0	4	40	374	175	36	3	632
7	3	3	68	462	216	28	0	780
8	1	6	49	306	140	28	2	532
Высокий	3	12	42	262	115	31	4	469
Всего	7	31	275	2576	1525	459	33	4906

Решение.

Вычислим коэффициент квадратической сопряженности $\chi^2=4906$

$$\left(\frac{5^2}{275\cdot192}+\frac{62^2}{2576\cdot192}+\ldots+\frac{8^2}{33\cdot192}+\frac{3^2}{275\cdot265}+\ldots+\frac{4^2}{33\cdot469}-1\right)=605{,}64.$$

Поскольку χ^2 существенно больше нуля, то связь между показателями можно считать доказанной. Проверим статистическую гипотезу $H_0:\chi^2=0$. По таблице хи-квадрат распределения находим $\chi^2_{0{,}05}(48)=65{,}16$ и, следовательно, $\chi^2>\chi^2_{0{,}05}$, т.е. между уровнем профессионализма и социальным слоем, к которому относят себя индивиды, действительно существует статистическая зависимость.

Коэффициент Крамера: $c=\left(\frac{\chi^2}{n\min(m_1-1,m_2-1)}\right)^{1/2}=$

$$=\left(\frac{605{,}64}{4906\cdot6}\right)^{1/2}=0{,}14.$$

Информационная мера связи $Y^2=2\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}n_{ij}\ln\left(\frac{n_{ij}}{n_{i\cdot}n_{\cdot j}/n}\right)=$

$=533{,}13.$

Показатель Пирсона $C=\sqrt{\frac{\chi^2}{n+\chi^2}}=\sqrt{\frac{605{,}64}{605{,}64+4906}}=0{,}33.$

Показатель связи Чупрова: $T=\sqrt{\frac{\chi^2}{n\sqrt{(m_1-1)(m_2-1)}}}=0{,}13,$

который, как видно, строже оценивает тесноту связи, чем показатель Пирсона.

Пример 3.4. В результате обследования работников предприятия получены следующие данные (чел.) Требуется оценить тесноту взаимосвязи между уровнем образования и удовлетворенностью своей работой с помощью коэффициентов контингенции и ассоциации.

Образование	Удовлетворены работой	Не удовлетворены работой	Итого
Высшее и среднее	300	50	350
Незаконченное среднее	200	250	450
Итого	500	300	800

Решение.

Коэффициент ассоциации

$Q=\frac{300\cdot250-50\cdot200}{\sqrt{350\cdot300\cdot500\cdot450}}=\frac{65000}{153707}=0{,}42$ и коэффициент контингенции — $Q_к=\frac{300\cdot250-50\cdot200}{300\cdot250+50\cdot200}=\frac{65000}{85000}=0{,}76.$

Значение коэффициента контингенции выше, чем коэффициента ассоциации. В целом оба показателя свидетельствуют об имеющейся связи между признаками.

Пример 3.5. Исследуется зависимость между стоимостью грузовой автомобильной перевозки Y (тыс. руб.), весом груза X_1 (тонн) и расстоянием X_2 (тыс. км) по 20 транспортным компаниям. Исходные данные приведены в таблице.

Y	51	16	74	7,5	33,0	26,0	11,5	52	15,8	8,0
X_1	35	16	18	2,0	14,0	33,0	20	25	13	2,0
X_2	2	1,1	2,55	1,7	2,4	1,55	0,6	2,3	1,4	2,1

Y	26	6,0	5,8	13,8	6,20	7,9	5,4	56,0	25,5	7,1
X_1	21	11,0	3	3,5	2,80	17,0	3,4	24,0	9,0	4,5
X_2	1,3	0,35	1,65	2,9	0,75	0,6	0,9	2,5	2,2	0,95

Решение.

В данном примере мы располагаем пространственной выборкой объема n=20, число объясняющих переменных p=2. Модель специфицируем в виде линейной функции: $Y=\beta_0+\beta_1X_1+\beta_2X_2+u$.

Следовательно, система нормальных уравнений будет иметь вид:

$$\begin{cases}\sum\limits_{i=1}^{n}Y_i=n\hat{\beta}_0+\hat{\beta}_1\sum\limits_{i=1}^{n}X_{1i}+\hat{\beta}_2\sum\limits_{i=1}^{n}X_{2i};\\ \sum\limits_{i=1}^{n}Y_iX_{1i}=\hat{\beta}_0\sum\limits_{i=1}^{n}X_{1i}+\hat{\beta}_1\sum\limits_{i=1}^{n}X_{1i}^2+\hat{\beta}_2\sum\limits_{i=1}^{n}X_{1i}X_{2i};\\ \sum\limits_{i=1}^{n}Y_iX_{2i}=\hat{\beta}_0\sum\limits_{i=1}^{n}X_{2i}+\hat{\beta}_1\sum\limits_{i=1}^{n}X_{2i}X_{1i}+\hat{\beta}_2\sum\limits_{i=1}^{n}X_{2i}^2.\end{cases}$$

Рассчитаем по данным таблицы необходимые для составления указанной системы суммы:

ΣY=454,5;	ΣX_1=277,2;	ΣX_2=31,8;
ΣY^2=18206,89;	ΣX_1^2=5860,9;	ΣX_2^2=61,45;
$\bar{Y}$=22,73;	$\bar{X}_1$=13,86;	$\bar{X}_2$=1,59;
ΣX_1Y=8912,57;	ΣX_2Y=908,56;	ΣX_1X_2=459,24;

Получим систему нормальных уравнений в виде:

$$\begin{cases}454,5=20\hat{\beta}_0+277,2\hat{\beta}_1+31,8\hat{\beta}_2;\\ 8912,57=277,2\hat{\beta}_0+5860,9\hat{\beta}_1+459,24\hat{\beta}_2;\\ 908,56=31,8\hat{\beta}_0+459,24\hat{\beta}_1+61,45\hat{\beta}_2.\end{cases}$$

Решая последнюю систему линейных алгебраических уравнений, например, методом Крамера, получим:

$\hat{\beta}_0$=-17,31; $\hat{\beta}_1$=1,16; $\hat{\beta}_2$=15,10.

Уравнение регрессии имеет вид: Y=-17,31+1,16·X_1+15,10·X_2.

Или, с учетом (3.17) и расчетов:

$$\sigma_Y=\sqrt{(18206,89-(454,5)^2/20)/20}=19,85;$$

$$\sigma_{X_1}=\sqrt{\left(\sum X_1^2-\frac{1}{n}\left(\sum X_1\right)^2\right)\Big/n}=\sqrt{(5860,9-(277,2)^2/20)/20}=10,05;$$

$$\sigma_{X_2}=\sqrt{\left(\sum X_2^2-\frac{1}{n}\left(\sum X_2\right)^2\right)\Big/n}=\sqrt{(61,45-(31,8)^2/20)/20}=0,74;$$

$$b_1=\beta_1\frac{\sigma_{X_1}}{\sigma_Y}=1,16\frac{10,05}{19,85}=0,77;\ b_2=\beta_2\frac{\sigma_{X_2}}{\sigma_Y}=15,10\frac{0,74}{19,85}=0,56.$$

Уравнение регрессии в стандартизованном масштабе:

$t_Y=0,77t_{X_1}+0,56t_{X_2}$.

То есть с ростом веса груза на одну сигму при неизменном расстоянии стоимость грузовых автомобильных перевозок увеличивается в среднем на 0,77 сигмы. Поскольку 0,77>0,56, то влияние веса груза на стоимость грузовых автомобильных перевозок больше, чем фактора расстояния.

Рассчитаем коэффициенты эластичности

$$\bar{Э}_{YX_1}=f'(\bar{X}_1)\frac{\bar{X}_1}{\bar{Y}}=\beta_1\frac{\bar{X}_1}{\beta_0+\beta_1\bar{X}_1+\beta_2\bar{X}_2}=1,16\cdot13,86/(-17,31+1,16\cdot13,86+$$

$$+15,10\cdot1,59)=0,71,\ \bar{Э}_{YX_1}=f'(\bar{X}_2)\frac{\bar{X}_2}{\bar{Y}}=\beta_2\frac{\bar{X}_2}{\beta_0+\beta_1\bar{X}_1+\beta_2\bar{X}_2}=1,05.$$

С увеличением среднего веса груза на 1% от его среднего уровня средняя стоимость перевозок возрастет на 0,71% от своего среднего уровня, при увеличении среднего расстояния

перевозок на 1% средняя стоимость доставки груза увеличится на 1,05%. Различия в силе влияния факторов на результат, полученные при сравнении уравнения регрессии в стандартизованном масштабе и коэффициентов эластичности, объясняются тем, что коэффициент эластичности рассчитывается исходя из соотношения средних, а стандартизованные коэффициенты регрессии из соотношения средних квадратических отклонений. Поскольку обычно статистики используют показатель грузооборота, вычисляемый как сумма произведений массы перевезенных грузов на расстояние перевозки, то построим регрессию стоимости 1 км грузовых автомобильных перевозок Y на грузооборот Q ($Q=X_1X_2$):

$$P = 5{,}88 + 0{,}48 \cdot Q - 0{,}003 \cdot Q^2,$$

причем регрессор $Q^2 = Q \cdot Q$ включен исходя из соображений известного экономического закона убывающей предельной полезности, согласно которому в данном случае стоимость перевозки на 1 км должна уменьшаться с ростом грузооборота, т.е. коэффициент при Q^2 должен иметь (и в построенном уравнении имеет) отрицательный знак.

Рассчитаем множественный коэффициент корреляции:

$$R_{YX_1X_2} = \sqrt{\frac{(0{,}6553)^2 + (0{,}6346)^2 - 2 \cdot 0{,}6553 \cdot 0{,}6346 \cdot 0{,}1247}{1 - (0{,}1247)^2}} = 0{,}860.$$

Величина множественного коэффициента корреляции, равного 0,860, свидетельствует о сильной взаимосвязи стоимости перевозки с весом груза и расстоянием, на которое он перевозится.

Коэффициент детерминации равен: $R^2=0{,}7399$. Скорректированный коэффициент детерминации рассчитываем по формуле (3.19): $R^2_{\text{скорр}} = 1-(1-0{,}7399)\cdot\frac{20-1}{20-2-1} = 0{,}709$.

Заметим, что величина скорректированного коэффициента детерминации отличается от величины коэффициента детерминации.

Таким образом, 70,9% вариации зависимой переменной (стоимости перевозки) объясняется вариацией независимых

переменных (весом груза и расстоянием перевозки). Остальные 29,1% вариации зависимой переменной объясняются факторами, не учтенными в модели.

Величина скорректированного коэффициента детерминации достаточно велика, следовательно, мы смогли учесть в модели наиболее существенные факторы, определяющие стоимость перевозки.

Заметим сначала, что в нашем примере:

$$\sum_i x_{1i}^2 = \sum_i X_{1i}^2 - n\bar{X}_1^2 = 5860{,}9 - 20 \cdot 13{,}86^2 = 2018{,}91;$$

$$\sum_i x_{2i}^2 = \sum_i X_{2i}^2 - n\bar{X}_2^2 = 61{,}45 - 20 \cdot 1{,}59^2 = 10{,}89;$$

$$\sum_i x_{1i}x_{2i} = \sum_i X_{1i}X_{2i} - n\bar{X}_1\bar{X}_2 = 459{,}24 - 20 \cdot 13{,}86 \cdot 1{,}59 = 18{,}49.$$

И матрица $\mathbf{X}^T\mathbf{X}$ имеет вид: $\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 2018{,}91 & 18{,}49 \\ 18{,}49 & 10{,}89 \end{bmatrix}$. Тогда обратная матрица будет $(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0{,}00050 & -0{,}00085 \\ -0{,}00085 & 0{,}09328 \end{bmatrix}$. Также $\mathbf{y}^T\mathbf{y} = \sum_i y_i^2 = \sum_i Y_i^2 - n\bar{Y}^2 = 18206{,}89 - 20 \cdot 22{,}73^2 = 7873{,}83$. Поэтому $\mathbf{e}^T\mathbf{e} = \mathbf{y}^T\mathbf{y}(1-R^2) = 7873{,}83 \cdot (1-0{,}74) = 2049{,}54$.

Проверим, например, гипотезу H_0: $\beta_1 = 0$. По (3.20) получим:

$$t = \frac{\hat{\beta}_j - \beta_{j0}}{\mu_{\hat{\beta}_j}} = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{q_{jj}}\sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2}} = 4{,}72.$$

Критическое значение для $\alpha=0{,}05$ будет равно $t_{0{,}025}(17)=2{,}110$, что меньше, чем расчетное 4,72, следовательно, гипотеза о равенстве нулю коэффициента β_1 отвергается.

Заполним таблицу дисперсионного анализа. Получаем $F = \frac{0{,}74}{2} : \frac{1-0{,}74}{20-2-1} = 24{,}17$, $F_{\alpha(2,17)} = 3{,}59$. В нашем примере $F > F_\alpha$, следовательно, нулевая гипотеза отклоняется, и уравнение множественной регрессии значимо.

Таблица дисперсионного анализа

Источник вариации	Сумма квадратов отклонений	Число степеней свободы	Средняя сумма квадратов отклонений
X_1, X_2	5828,84	2	2914,42
Остаток	2049,54	17	120,56
Общая вариация	7878,38	19	

Проверим, например, гипотезу $H_0: \beta_2 - 10\beta_1 = 3$. В этом случае $\mathbf{c}^T = (-10, 1)$ и $\theta = 3$. Вычислим t-статистику по формуле (3.24):

$$t = \frac{\mathbf{c}^T\hat{\boldsymbol{\beta}} - \theta}{\sqrt{\hat{\sigma}^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}} = \frac{-1{,}16 \cdot 10 + 15{,}1 - 3}{\sqrt{2049{,}54 \cdot (100 \cdot 0{,}0005 - 2 \cdot 10 \cdot (-0{,}00085) + 0{,}09328)}} = 0{,}028,$$

что меньше критического $t_{0,05}(17) = 1{,}740$ и, следовательно, гипотеза H_0 принимается.

Вопросы и задачи

1. Имеется корреляционная матрица признаков X_1 — доход за месяц, X_2 — число лет образования, X_3 — процент времени, который занимает основная работа $\mathbf{R} = \begin{bmatrix} 1 & 0{,}24 & 0{,}01 \\ 0{,}24 & 1 & 0{,}09 \\ 0{,}01 & 0{,}09 & 1 \end{bmatrix}$, вычисленная по массиву данных msm.sta (число наблюдений 2874). Рассчитайте парные и множественные (в том числе частные) коэффициенты корреляции.

2. Кратко поясните особенности множественного коэффициента корреляции, частного коэффициента корреляции.

3. Два эксперта проранжировали 10 предложенных им проектов по степени эффективности: $X_1 = (1; 2; 4; 6; 7; 3; 5; 8; 9; 10)$, $X_2 = (2; 3; 1; 4; 6; 5; 9; 7; 8; 10)$. Оцените степень согласованности мнений экспертов, вычислив ранговые коэффициенты корреляции Спирмена и Кендалла.

4. Имеются результаты опроса группы из 10 экспертов по трем вопросам социально-экономической политики:

Значение экспертной оценки по вопросам	Порядковый номер эксперта									
	1	2	3	4	5	6	7	8	9	10
1	0,35	0,52	0,48	0,64	0,69	0,2	0,36	0,65	0,78	0,54
2	30	35	38	40	45	31	33	46	42	37
3	2,5	3,1	3	3,6	4,5	2,8	3,7	4,2	4,7	2,9

Требуется оценить согласованность мнений экспертов с помощью коэффициента конкордации и проверить его значимость на уровне 5 %. Сделать вывод.

5. В чем особенности измерения степени тесноты статистической связи между категоризованными переменными?

6. Проанализируйте связь между полом работника и характером труда в сезонных отраслях:

Пол	Численность занятых в отраслях		
	Сезонных	Не сезонных	Всего
Мужчины	187	265	452
Женщины	307	272	579
Всего	494	537	1031

7. Оценка студентами профессиональных качеств преподавателей представлена в таблице ниже. Рассчитайте коэффициент взаимной сопряженности Пирсона, Крамера.

Критерии оценки качества преподавателей	Оценка				
	Высокая	Средняя	Низкая	Затрудняюсь ответить	Итого
Знание предмета	62	26	1	11	100
Умение обучать	21	61	8	10	100
Восприимчиво-сть к новому	20	51	10	19	100
Способность к саморазвитию	25	51	10	14	100
Итого	128	189	29	54	400

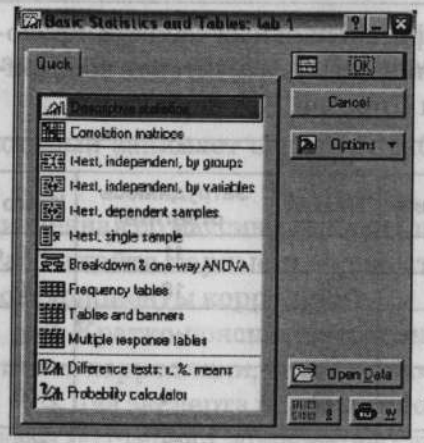
8. Имеется распределение двух переменных X и Y. Известно: $f(y=2|x=1) = 0{,}3927$, $f(x=1) = 0{,}2865$, $f(x=2) = 0{,}265$, $f(y=4) = 0{,}2475$ и $E[x|y=2] = 1{,}9262$.

Заполните в таблице недостающие значения. Вычислите по таблице: 1) $E(X)$, $V(X)$, 2) $E(Y)$, $V(Y)$, 3) $V(X|Y=2)$, 4) $r_{X,Y}$.

X, Y	1	2	3	4	Итого
1	260			370	
2	320		320		1060
3			290	280	911
4	250		275	300	
Итого					1

Задание к лабораторному практикуму
Время выполнения — 2 часа

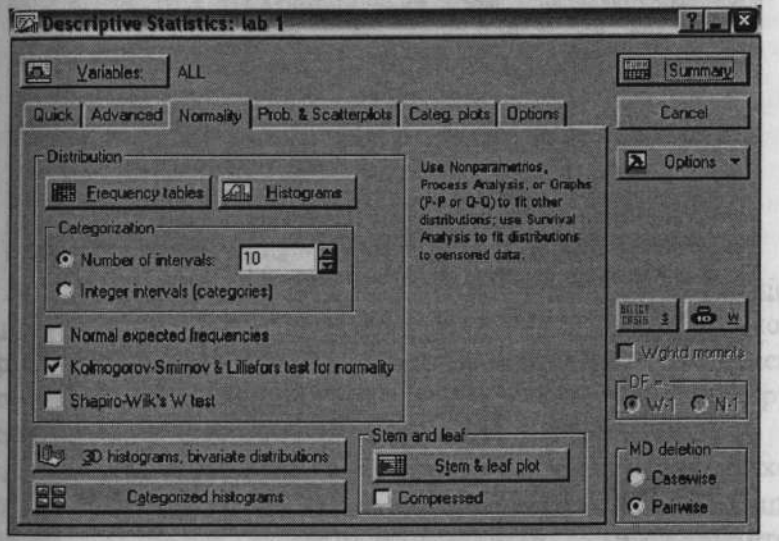
1. Запустите приложение *Statistica*. Откройте файл *Lab1.sta*, воспользовавшись меню File\Open. В окне переменных вы увидите четыре переменные. Файл содержит подвыборку 4794 наблюдений из массива msm.sta по индивидам. Описание переменных: *lw* — логарифм заработной платы, *edu* — число лет образования, *expr* — опыт работы, *expr2* — квадрат переменной *expr*.



2. Просмотрите дескриптивные статистики переменных в выборке: Statistics\Basic Statistics and Tables\Descriptive statistics. Нажав вкладку Variables, задайте переменные, выбрав все переменные. Воспользовавшись вкладкой Advanced, поставьте флажки для вывода медианы (Median), моды (Mode), дисперсии (Variance), эксцесса (Kurtosis) и асимметрии (Skewness). Нажмите кнопку Summary и получите таблицу результатов. Сделайте выводы.

Вернитесь к диалоговому окну Descriptive statistics и, выбрав вкладку Normality, нажмите кнопку Histograms. Просмотрите полученные диаграммы распределения для каждой из переменных, а также сравните их с кривой нормального распределения. Вверху каждой гистограммы приведено значение статистики Колмогорова-Смирнова для проверки соответствия эмпирического распределения нормальному закону распределения. Сделайте выводы об эмпирическом распределении каждой переменной.

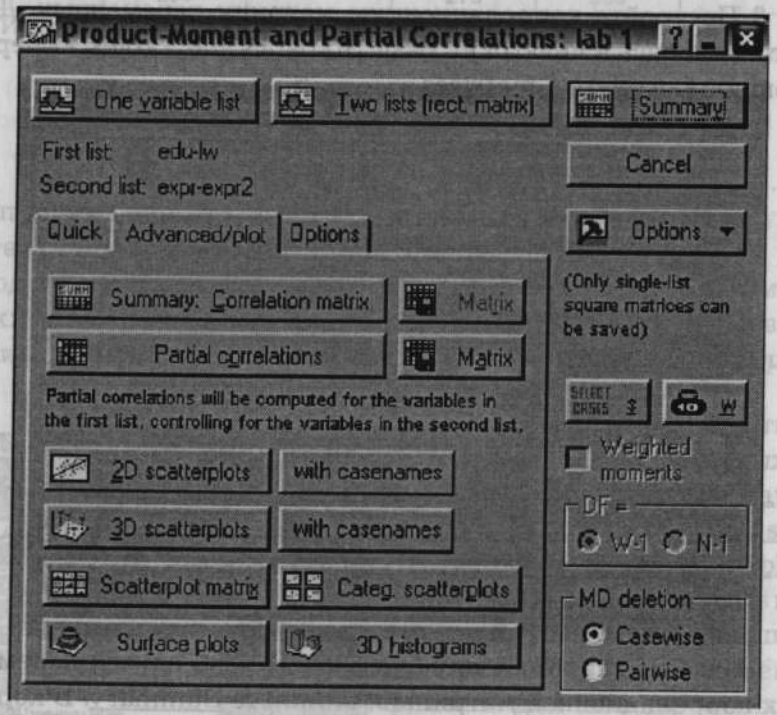
3. Постройте корреляционную матрицу переменных. Для этого, находясь в меню окна Basic Statistics and Tables, выберите опцию Correlation matrices.



Задайте, нажав кнопку One variable list, список переменных корреляционной матрицы и кликнете Summary. В корреляционной матрице значимые (по умолчанию на 5% уровне) коэффициенты будут выделены красным цветом. Прокомментируйте результаты.

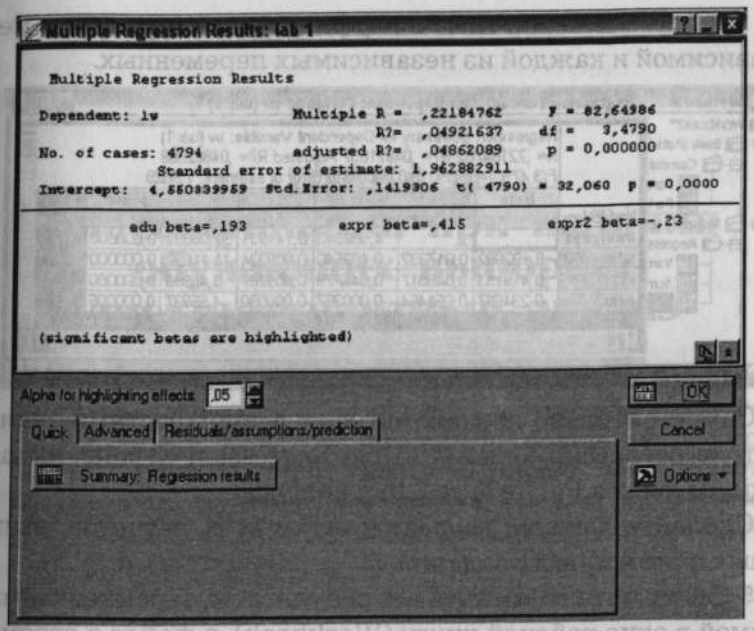
Также можно рассчитать частные коэффициенты корреляции. Например, для расчета частного коэффициента корреляции между переменными *edu* и *lw* при фиксированном значении переменных *expr* и *expr2* необходимо с помощью кнопки Two lists задать первый лист (слева) из переменных *edu* и *lw* и

затем второй лист (справа) из переменных *expr* и *expr2*. Затем в окне Product-Moment and Partial correlation необходимо выбрать вкладку Advanced/plot и в появившемся окне — опцию Partial correlations. Получим частный коэффициент корреляции. Аналогично получите остальные частные коэффициенты корреляции. Сравните их с парными коэффициентами корреляции и сделайте выводы.



4. Постройте уравнение линейной множественной регрессии переменной заработной платы *lw* от переменных *edu*, *expr* и *expr2*.

Для этого откройте меню Statistics\Multiple Regression и, выбрав опцию Variables, укажите в появившемся окне слева зависимую переменную (*lw*), в правом окне независимые переменные (*edu*, *expr* и *expr2*). Нажмите OK.

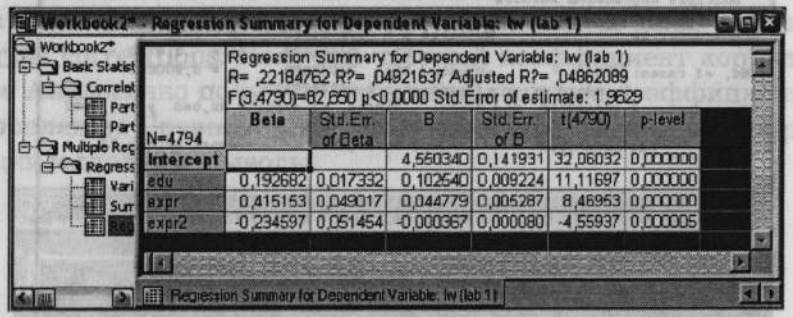


Появится окно результатов регрессии Multiple Regression Results. В окне результатов имеется значение множественного коэффициента корреляции зависимой переменной с независимыми, значение коэффициента детерминации, значения других статистик, относящихся к уравнению регрессии.

Для просмотра коэффициентов регрессии и относящихся к ним статистик нажмите в окне результатов на кнопку Summary: Regression results. Коэффициенты регрессии находятся в столбце с именем B, затем следует столбец стандартных ошибок коэффициентов и значения соответствующих t-статистик. Значимые коэффициенты регрессии выделены красным цветом. Вернитесь в окно Multiple Regression Results и выберите вкладку Advanced и затем опцию ANOVA. В появившемся окне результатов представлена таблица дисперсионного анализа для уравнения регрессии. В случае значимости уравнения в целом значения F-критерия будут выделены красным цветом.

Далее в окне Multiple Regression Results на вкладке Advanced выберите опцию Partial correlations. В появившемся

окне представлены частные коэффициенты корреляции между зависимой и каждой из независимых переменных.



Самостоятельно познакомьтесь с возможностями анализа остатков регрессии (окно Multiple Regression Results, вкладка Residuals, опция Perform residual analysis).

Сделайте выводы по результатам всех расчетов, выполненных в этом пункте задания.

5. Вы можете сохранить все результаты, запомненные программой в окне рабочей книги (Workbook), в файле в своем рабочем каталоге.

ГЛАВА 4
Классификация при наличии обучающих выборок: дискриминантный анализ

Всякий необходимо причиняет пользу, употреблённый на своём месте. Напротив того: упражнения лучшего танцмейстера в химии неуместны; советы опытного астронома в танцах глупы.

К. Прутков. Мысли и афоризмы

4.1. Основные определения

При наблюдении больших статистических совокупностей часто появляется необходимость разделить неоднородную совокупность на однородные группы (классы). Такое расчленение позволяет получить лучшие результаты при дальнейшем моделировании зависимостей между отдельными признаками. Например, разбиение предприятий на несколько однородных групп по показателям производственно-хозяйственной деятельности; подбор кандидатов на определенную должность; оценка кредитоспособности заемщика в банковской деятельности и т.п.

Классификация — разделение рассматриваемой совокупности объектов или явлений на однородные в определенном смысле группы, либо отнесение каждого из заданного множества объектов к одному из заранее известных классов.

Методы дискриминантного анализа разрабатывались П.Ч. Махаланобисом, Р. Фишером, Г. Хотеллингом и др.

Пусть каждый из n объектов характеризуется p свойствами, т.е. имеем многомерные наблюдения X_i, такие что $X_i=(x_i^{(1)},x_i^{(2)},...,x_i^{(p)}), i=1,...n$. Фактически задана матрица вида (1.1) или (1.2).

На «входе» задачи классификации мы имеем:

а) априорные сведения о классах: число классов, общий вид или свойства законов распределения X внутри каждого из классов, диапазон изменения анализируемых показателей;

б) где обучающие выборки X_{j1}, X_{j2}..., X_{jn_j}, $j=1,2,...,k$, k — число классов.

На «выходе» задачи классификации:

а) набор наиболее информативных объясняющих переменных (типообразующих признаков), которые либо отбираются по определенному правилу из числа исходных признаков $x^{(1)},...,x^{(p)}$, либо строятся в качестве некоторых их комбинаций.

б) правило отнесения (дискриминантная функция, классификатор) каждого классифицируемого объекта, заданного значениями описательных признаков X_i к одному из классов.

Если на «входе» есть обучающие выборки, то возникает задача *классификации с обучением*, если нет обучающих выборок — *классификации без обучения*.

Класс — генеральная совокупность, описываемая одномодальной функцией плотности $f(x)$.

Дискриминантная функция (решающее правило, процедура классификации) — статистика, служащая для построения правила классификации объектов по группам.

На рис. 4.1 показаны два множества S_1 и S_2 объектов, характеризующихся двумя признаками x_1 и x_2. По каждой переменной отдельно объекты разных множеств имеют сходные характеристики — это видно по проекциям объектов на каждую ось. Чтобы наилучшим образом разделить два множества, нужно построить линейную комбинацию переменных x_1 и x_2: $D(x)=b_1x_1+b_2x_2$. Функция $D(X)$ называется *канонической дискриминантной функцией*, а величины b_1 и b_2 — *коэффициентами дискриминантной функции*. Фактически для двумерного случая это означает определение новой системы координат с осями L и C, причем проекции объектов разных множеств на ось L должны быть максимально разделены [9].

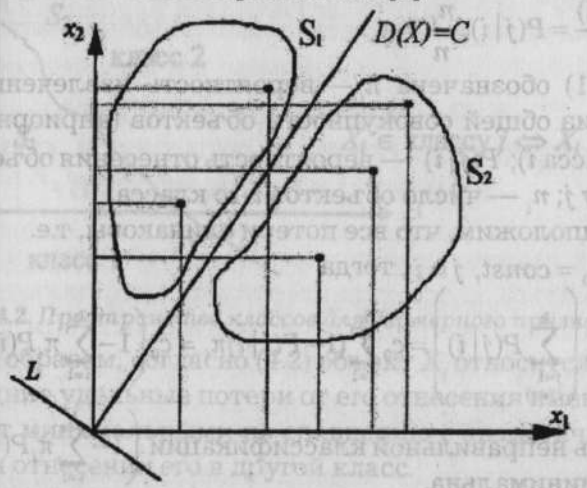


Рис. 4.1. *Геометрическая интерпретация дискриминантной функции*

Дискриминантная функция может быть как линейной, так и нелинейной. Выбор ее вида зависит от геометрического расположения разделяемых классов в пространстве дискриминантных переменных.

В основу вероятностных методов классификации положен *принцип*: наблюдение относится к тому классу, в рамках которого оно выглядит более правдоподобным. Принцип корректируется с учетом цели минимизации потерь от неправильной классификации [1].

Пусть $c(j|i)$ — функция потерь при отнесении объекта i-го класса к классу j; $m(j|i)$ — число объектов i-го класса неправильно классифицированных в класс j. Тогда общие потери составят $c_n=\sum_{i=1}^{k}\sum_{j=1}^{k}c(j|i)m(j|i)$, где k — число классов.

Пусть n — число классифицируемых объектов. Тогда в пределе по n:

$$c=\lim_{n\to\infty}\left(\frac{1}{n}c_n\right)=\lim_{n\to\infty}\sum_i\sum_j c(j|i)\frac{m(j|i)}{n_i}\frac{n_i}{n}=\sum_i\pi_i\sum_j c(j|i)P(j|i), \quad (4.1)$$

где $\frac{m(j|i)}{n_i}=P(j|i); \frac{n_i}{n}=\pi_i$.

В (4.1) обозначена π_i — вероятность извлечения объекта класса i из общей совокупности объектов (априорная вероятность класса i); $P(j|i)$ — вероятность отнесения объекта класса i к классу j; n_i — число объектов i-го класса.

Предположим, что все потери одинаковы, т.е. $c(j|i)=c_0=const$, $j\neq i$, тогда

$c=c_0\sum_{i=1}^{k}\pi_i\left(\sum_{\substack{j=1\\(j\neq i)}}^{k}P(j|i)\right)=c_0\sum_{i=1}^{k}(1-P(i|i)\pi_i=c_0\left(1-\sum_{i=1}^{k}\pi_iP(i|i)\right)$ и вероятность неправильной классификации $\left(1-\sum_{i=1}^{k}\pi_iP(i|i)\right)$ должна быть минимальна.

Пусть генеральная совокупность описывается смесью k классов с плотностью вероятности $f(X)=\sum_{j=1}^{k}\pi_jf_j(X)$.

Обозначим $\delta(X)$ дискриминантную функцию, которая принимает значения 1,2,...,k, причем, если $\delta(X)=j$, то наблюдение X относится к классу j: $S_j=\{X:\delta(X)=j\}, j=1,...,k$, *где* S_j — p-мерные области в пространстве возможных значений многомерного признака X (рис. 4.2).

Дискриминантная функция $\delta(X)$ называется *оптимальной* (байесовской), если она сопровождается минимальными потерями (4.1) среди всех других процедур классификации.

Можно показать [1], что оптимальное байесовское правило классификации (ОБПК) имеет вид:

$$S_j^{onm}=\left\{X:\sum_{i=1,i\neq j}^{k}\pi_if_i(X)c(j|i)=\min_{1\le l\le k}\sum_{i=1,i\neq l}^{k}\pi_if_i(X)c(l|i)\right\}. \quad (4.2)$$

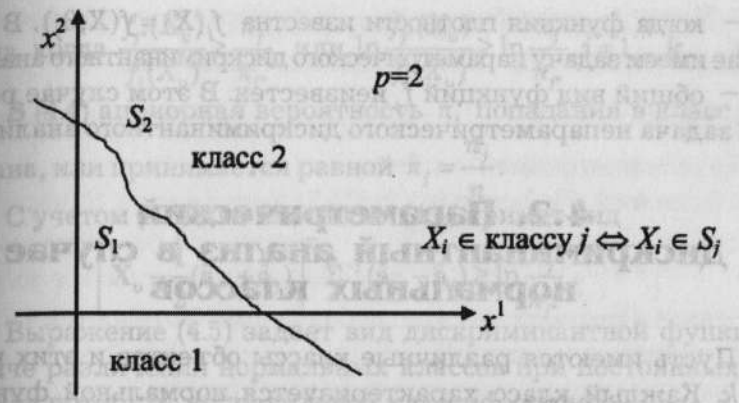


Рис. 4.2. *Пространство классов для двумерного признака*

Таким образом, согласно (4.2) объект X_j относится к классу j, если средние удельные потери от его отнесения именно в этот класс будут минимальными по сравнению с аналогичными потерями при отнесении его в другой класс.

Графически потери от неправильной классификации проиллюстрированы на рис. 4.3, и потери от неправильной классификации в этом случае будут $c=c(2|1)P(2|1)\pi_1+c(1|2)P(1|2)\pi_2$.

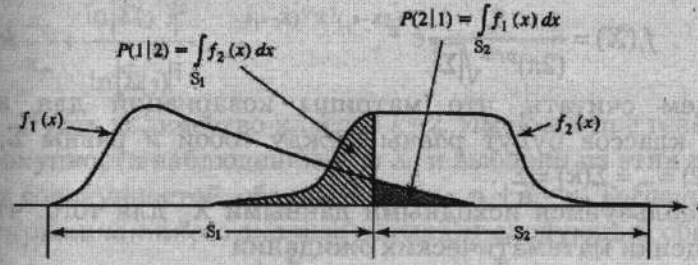


Рис. 4.3. *Неправильная классификация для двух гипотетических классов*

Если $c(j|i)=c_0=const$, то в (4.2) $\pi_jf_j(X_j)=\max_{1\le l\le k}\pi_lf_l(X_l)$. Но в (4.2) надо знать π_j и f_j. Заметим, что по выборке можно получить оценки $\hat{\pi}_j=\frac{n_j}{n}$, где n_j — объем j-й обучающей выборки, n — общее количество объектов. Для оценки f_j возможны две ситуации:

– когда функция плотности известна $f_j(X)=f(X,\theta_j)$. В этом случае имеем задачу параметрического дискриминантного анализа;

– общий вид функций f_j неизвестен. В этом случае решается задача непараметрического дискриминантного анализа.

4.2. Параметрический дискриминантный анализ в случае нормальных классов

Пусть имеются различные классы объектов и этих классов k. Каждый класс характеризуется нормальной функцией плотности, зависящей от p переменных. Предполагается, что задана обучающая выборка $X_{j1}, X_{j2}, \ldots, X_{jn_j}$, где $j=1,2,\ldots,k$, $X_{ji}=(x_{ji}^{(1)},x_{ji}^{(2)},\ldots,x_{ji}^{(p)})$, $i=1,\ldots,n_j$, здесь n_j — количество наблюдений в классе j, j — индекс по классам, i — индекс по наблюдениям внутри класса, l — индекс по переменным $l=1,\ldots,p$.

Пусть a_j — математическое ожидание $a_j=E(X_j)$ случайной величины, относящейся к j-му классу, Σ — матрица ковариаций, причем

$$f_j(X)=\frac{1}{(2\pi)^{p/2}\sqrt{|\Sigma|}}e^{-\frac{1}{2}(X-a_j)^T\Sigma^{-1}(X-a_j)}. \qquad (4.3)$$

Будем считать, что матрицы ковариаций для каждого из классов будут *равны* между собой и равны Σ, т.е. $\Sigma(1)=\Sigma(2)=\ldots=\Sigma(k)=\Sigma$.

Воспользуемся исходными данными X_{ji} для того, чтобы найти оценки математических ожиданий

$$\hat{a}_j^{(l)}=\frac{1}{n_j}\sum_{i=1}^{n_j}x_{ji}^{(l)}, l=1,..,p, j=1,..,k \text{ и ковариаций } \hat{\Sigma}=\left\{\hat{\sigma}_{lq}\right\}_{l,q=1}^{p};$$

$$\hat{\sigma}_{lq}=\frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ji}^{(l)}-\hat{a}_j^{(l)}\right)\left(x_{ji}^{(q)}-\hat{a}_j^{(q)}\right), l,q=1,\ldots,p, n=n_1+n_2+\ldots+n_k.$$

Тогда с учетом нормального распределения каждого из классов правило классификации (4.2) примет следующий вид: наблюдение X_v относится к классу с номером j^* тогда и только

тогда, когда $\frac{f_{j^*}(X_v)}{f_j(X_v)}\geq\frac{\pi_j}{\pi_{j^*}}$, или $\ln\frac{f_{j^*}(X_v)}{f_j(X_v)}\geq\ln\frac{\pi_j}{\pi_{j^*}}, j=1,\ldots,k$. (4.4.)

В (4.4) априорная вероятность π_j попадания в класс j или задана, или принимается равной $\hat{\pi}_j=\frac{n_j}{n}$.

С учетом (4.3) для плотности (4.4) примет вид

$$\left[\mathbf{X}_v-\frac{1}{2}(\hat{\mathbf{a}}_{j^*}+\hat{\mathbf{a}}_j)\right]^T\hat{\Sigma}^{-1}(\hat{\mathbf{a}}_{j^*}-\hat{\mathbf{a}}_j)\geq\ln\frac{\pi_j}{\pi_{j^*}}. \qquad (4.5)$$

Выражение (4.5) задает вид дискриминантной функции в задаче различения нормальных классов при постоянных значениях потерь от неправильной классификации.

Функция $W(\mathbf{X})=\left[\mathbf{X}_v-\frac{1}{2}(\hat{\mathbf{a}}_{j^*}+\hat{\mathbf{a}}_j)\right]^T\hat{\Sigma}^{-1}(\hat{\mathbf{a}}_{j^*}-\hat{\mathbf{a}}_j)$ называется *линейной дискриминантной функцией Фишера.*

Если предположение о равенстве ковариационных матриц не выполняется, $\Sigma(1)\neq\Sigma(2)\neq\ldots\neq\Sigma(k)$, то решающее правило примет вид: $-\frac{1}{2}\left[(\mathbf{X}_v-\hat{\mathbf{a}}_{j^*})^T\hat{\Sigma}^{-1}(j^*)(\mathbf{X}_v-\hat{\mathbf{a}}_{j^*})-(\mathbf{X}_v-\hat{\mathbf{a}}_j)^T\hat{\Sigma}^{-1}(j)(\mathbf{X}_v-\hat{\mathbf{a}}_j)\right]\geq$

$$\geq\ln\frac{\pi_j}{\pi_{j^*}}+\frac{\ln\left|\hat{\Sigma}(j^*)\right|^{\frac{1}{2}}}{\ln\left|\hat{\Sigma}(j)\right|^{\frac{1}{2}}}.$$

Пусть количество классов $k=2$. Имеются две генеральные совокупности наблюдений X_1 и X_2 и выборки из этих генеральных совокупностей, объемом n_1 и n_2, $n_1+n_2=n$. Тогда *алгоритм дискриминантного анализа* выглядит следующим образом:

1. Находим $\hat{a}_1^{(l)}=\frac{1}{n_1}\sum_{i=1}^{n_1}x_{1i}^{(l)}$, $\hat{a}_2^{(l)}=\frac{1}{n_2}\sum_{i=1}^{n_2}x_{2i}^{(l)}, l=1,\ldots,p$.

2. Рассчитываем $(\hat{\mathbf{a}}_1-\hat{\mathbf{a}}_2)$ и $\frac{1}{2}(\hat{\mathbf{a}}_1+\hat{\mathbf{a}}_2)$.

3. Определяем ковариационную матрицу

$$\hat{\sigma}_{lq}=\frac{1}{n_1+n_2-2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(x_{ji}^{(l)}-\hat{a}_j^{(l)})(x_{ji}^{(q)}-\hat{a}_j^{(q)}), \; l,q=1,\ldots,p, \; \hat{\Sigma}=\left\{\hat{\sigma}_{lq}\right\}.$$

4. Находим матрицу $\hat{\Sigma}^{-1}$, обратную к матрице $\hat{\Sigma}$

5. Находим *коэффициенты дискриминантной функции* $\mathbf{w} = \hat{\Sigma}^{-1}(\hat{\mathbf{a}}_1 - \hat{\mathbf{a}}_2)$.

6. Вычисляем $X_0 - \frac{1}{2}(\hat{\mathbf{a}}_1 + \hat{\mathbf{a}}_2)$, где X_0 — наблюдение, которое надо классифицировать.

7. Если $k=2$ и $\pi_1 = \pi_2 = 0{,}5$, то (4.5) имеет вид:

$$\left[X_0 - \frac{1}{2}(\hat{\mathbf{a}}_1 + \hat{\mathbf{a}}_2)\right]^T \hat{\Sigma}^{-1}(\hat{\mathbf{a}}_1 - \hat{\mathbf{a}}_2) \geq 0 . \qquad (4.6)$$

Если (4.6) выполняется, то наблюдение X_0 относится к классу 1, если нет, то к классу 2.

Графическая иллюстрация процедуры Фишера для двумерного случая представлена на рис. 4.4.

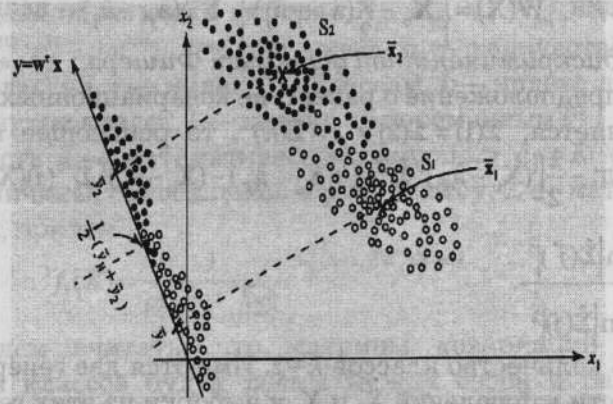


Рис. 4.4. *Классификация по Фишеру для двумерного пространства*

Для оценки вклада отдельной переменной в значение дискриминантной функции целесообразно пользоваться стандартизованными коэффициентами дискриминантной функции. Для этого исходные признаки x_{ji}^l стандартизируют вычитанием среднего и делением на дисперсию.

4.3. Непараметрический дискриминантный анализ

Применение дискриминантного анализа возможно также в случае, когда общий параметрический вид функций $f_j(X)$ неизвестен. В этом случае в (4.2) функции $f_j(X)$ заменяются непараметрическими оценками $\hat{f}_j(X)$, построенными по соответствующим обучающим выборкам.

В качестве непараметрических оценок $\hat{f}_j(X)$ чаще всего используют «ядерные» оценки, получаемые по формуле

$$\hat{f}_j(X) = \frac{1}{n_j h}\sum_{i=1}^{n_j} K\left(\frac{X - X_i}{h}\right),$$

где h — так называемая «ширина окна»; $K(z)$ — некоторая ядерная функция $\left(\frac{X - X_i}{h}\right) = z$, например, используются функции Парзена $K(z) = \begin{cases} \frac{4}{3} - 8z^2 + 8|z|^3, |z| \leq \frac{1}{2} \\ \frac{8(1-|z|)^3}{3}, \frac{1}{2} < |z| \leq 1; \\ 0, \text{ иные} \end{cases}$ Гаусса $K(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$;

Епанечникова $K(z) = \begin{cases} \frac{\frac{3}{4}\left(1 - \frac{1}{5}z^2\right)}{\sqrt{5}}, |z| < \sqrt{5} \text{ и другие.} \\ 0, \text{ иные} \end{cases}$

Существуют непараметрические методы дискриминантного анализа, не использующие прямых непараметрических оценок плотностей. Например, к таким методам относится метод N ближайших соседей (Фикс и Ходжес, 1951 г.). Суть метода: вокруг классифицируемой точки X_0 описывается сфера минимального радиуса (рис. 4.5), содержащая N элементов из обучающих выборок. Точку X_0 относят к тому классу, представителей которого в сфере оказалось больше, чем представителей любого другого класса. Как правило, величина $N \approx n^{\frac{4}{4+p}}$.

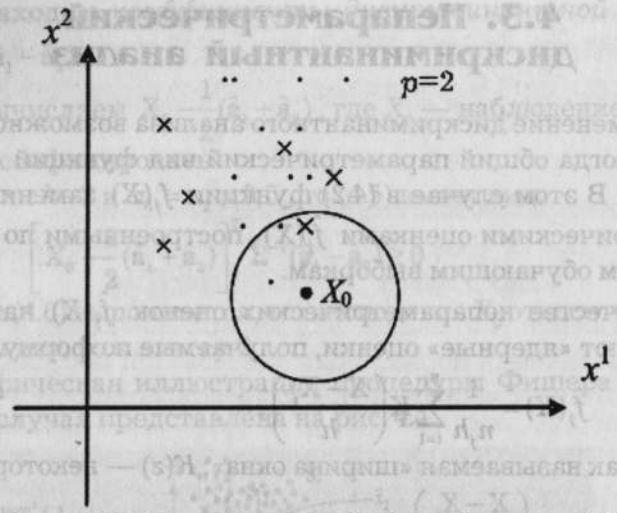


Рис. 4.5. *Иллюстрация метода ближайших соседей*

4.4. Оценка качества дискриминантной функции и информативности отдельных признаков

Дискриминантный анализ (ДА) выполняется в предположении, что объекты принадлежат одному из двух (или более) классов.

Существуют ограничения, касающиеся свойств дискриминантных переменных:

– количество дискриминантных переменных не должно превышать число объектов за вычетом двух, т.е. $0 < p < (n-2)$;

– переменные должны быть измерены в метрической шкале;

– дискриминантные переменные должны быть линейно независимы.

Также ДА выполняется в предположении, что обучающие выборки взяты из нормально распределенной совокупности. Нарушение этого предположения не является фатальным и

приводит к потерям в эффективности и точности при проверке значимости дискриминации. Более важно выполнение второго предположения о равенстве ковариационных матриц для всех классов, т.е. фактически о том, что все классы характеризуются одними и теми же внутригрупповыми дисперсиями и корреляциями.

Для проверки качества построенных дискриминантных функций, используется статистическая проверка гипотезы: $H_0 : a_1 = a_2 = ... = a_k$ о равенстве математических ожиданий внутри классов. Для проверки такой гипотезы может быть использовано обобщенное расстояние Махаланобиса: $D^2 = \sum_{j=1}^{k} n_j (\mathbf{a}_j - \mathbf{a})^T \mathbf{\Sigma}^{-1} \times (\mathbf{a}_j - \mathbf{a})$, представляющее собой взвешенную сумму расстояний от вектора средних каждого класса ($\mathbf{a}_j$) до общего вектора средних $\mathbf{a}$. Если нулевая гипотеза верна, то D^2 аппроксимируется F-распределением Фишера.

Другим способом проверки гипотезы H_0 является использование статистики Уилкса (лямбда Уилкса):

$$\lambda = \frac{|\Sigma|}{|\mathbf{T}|}, \text{ где } \hat{\mathbf{\Sigma}} = \left\{\hat{\sigma}_{lq}\right\}, \hat{\mathbf{T}} = \left\{\hat{t}_{lq}\right\}$$

$$\text{и } \hat{t}_{lq} = \frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji}^{(l)} - \hat{a}^{(l)})(x_{ji}^{(q)} - \hat{a}^{(q)}),$$

$$\hat{\sigma}_{lq} = \frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji}^{(l)} - \hat{a}_j^{(l)})(x_{ji}^{(q)} - \hat{a}_j^{(q)}), l, q = 1,2,...,p.$$

Используются оценки матрицы внутригрупповой ковариации $\mathbf{\Sigma}$ и полной ковариационной матрицы $\mathbf{T}$. Статистика Уилкса также аппроксимируется F-распределением. Если λ=0, то дискриминация считается идеальной, если λ=1, то она считается плохой.

Кроме задачи классификации наблюдений в один из классов, интерес представляет задача прогнозирования. Насколько хорошо построенные дискриминирующие функции могут предсказывать, к какому классу принадлежит конкретное наблюдение, анализируют таблицу сопряженности «Факт — Прогноз»

и оценивают вероятность ошибочной классификации каждого класса. Если процент ошибок высок, то причиной этого является либо нарушение предположений о нормальности многомерного распределения и равенстве ковариационных матриц, либо использование плохих дискриминантных переменных.

Также можно использовать следующий прием (так называемая «перекрестная проверка»): один из элементов обучающей выборки исключается и строится классификация по оставшимся элементам и удаленный элемент используется в качестве экзаменующего. Затем удаленный возвращается в выборку, удаляется второй элемент и процедура повторяется. Алгоритм заканчивает свою работу, когда удалению подвергнутся все элементы выборки по очереди. Ошибка дискриминации складывается из ошибок дискриминации каждого экзаменующего элемента.

Пошаговый дискриминантный анализ. Наиболее общим принципом применения дискриминантного анализа является включение в исследование по возможности большего числа переменных p, значимых для дискриминации. С целью определения тех из них, которые наилучшим образом разделяют выборки между собой, используется пошаговая процедура, в которой на каждом шаге построения модели дискриминации просматриваются все переменные и находится та из них, которая вносит наибольший вклад в различение между совокупностями, и эта переменная включается в модель на текущем шаге, происходит переход к следующему шагу.

Включение переменных осуществляется по критерию Фишера, которое указывает на статистическую значимость переменной для дискриминантного анализа. Аналогично из множества всех включенных переменных можно последовательно исключать незначимые для дискриминации, и таким образом из общего числа первоначально включенных p переменных остается p^* переменных ($p^* < p$), чей вклад в дискриминацию больше. В первом случае метод называется пошаговым включением, во втором — исключением.

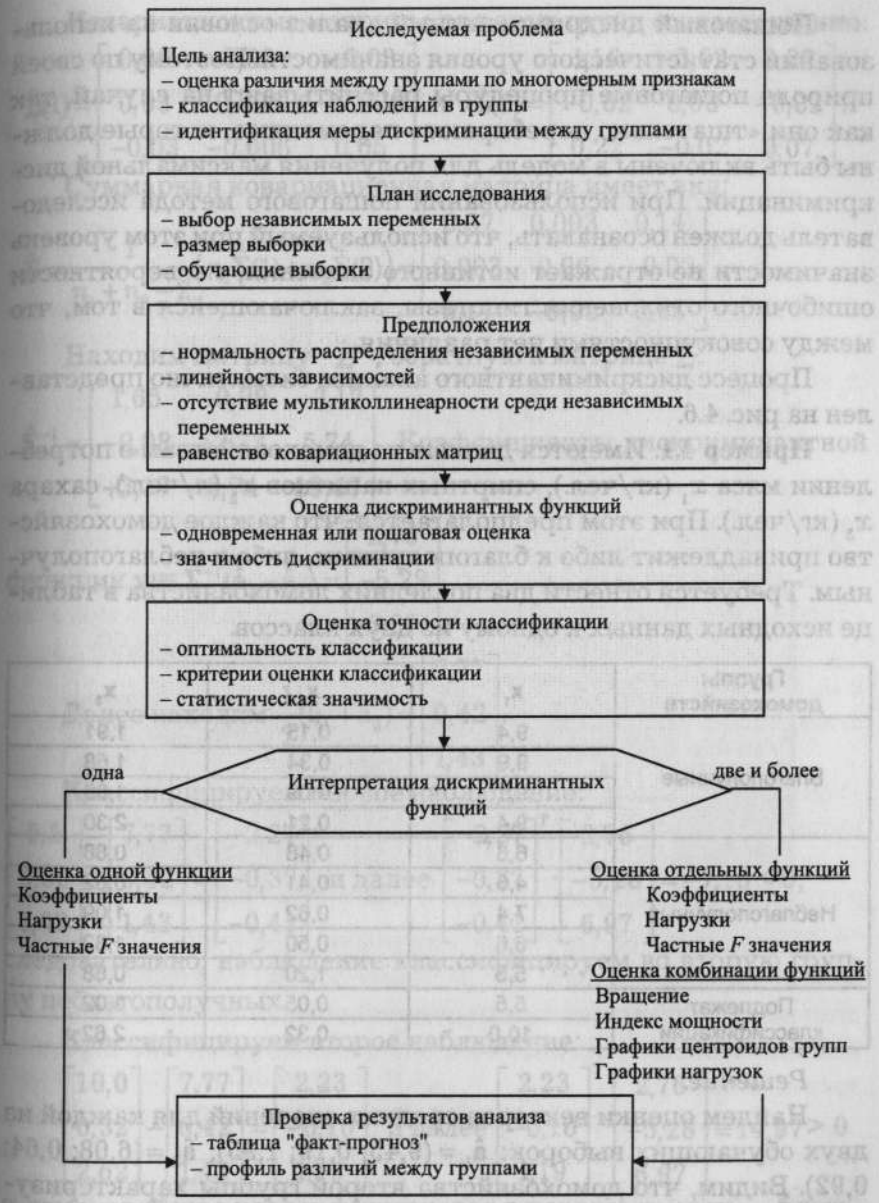


Рис. 4.6. *Блок-схема принятия решений в процессе дискриминантного анализа*

Пошаговый дискриминантный анализ основан на использовании статистического уровня значимости. Поэтому по своей природе пошаговые процедуры рассчитывают на случай, так как они «тщательно перебирают» переменные, которые должны быть включены в модель для получения максимальной дискриминации. При использовании пошагового метода исследователь должен осознавать, что используемый при этом уровень значимости не отражает истинного значения, т.е. вероятности ошибочного отклонения гипотезы, заключающейся в том, что между совокупностями нет различия.

Процесс дискриминантного анализа схематично представлен на рис. 4.6.

Пример 4.1. Имеются данные по домохозяйствам о потреблении мяса x_1 (кг/чел.), спиртных напитков x_2 (л/чел.), сахара x_3 (кг/чел.). При этом предполагается, что каждое домохозяйство принадлежит либо к благополучным, либо к неблагополучным. Требуется отнести два последних домохозяйства в таблице исходных данных к одному из двух классов.

Группы домохозяйств	x_1	x_2	x_3
Благополучные	9,4	0,15	1,91
	9,9	0,34	1,68
	9,1	0,09	1,89
	9,4	0,21	2,30
Неблагополучные	6,6	0,48	0,88
	4,3	0,41	0,62
	7,4	0,62	1,09
	6,6	0,50	1,32
	5,5	1,20	0,68
Подлежат классификации	5,5	0,05	1,02
	10,0	0,32	2,62

Решение.

Найдем оценки векторов средних значений для каждой из двух обучающих выборок: $\hat{\mathbf{a}}_1 = (9{,}45;\ 0{,}19;\ 1{,}95)$, $\hat{\mathbf{a}}_2 = (6{,}08;\ 0{,}64;\ 0{,}92)$. Видим, что домохозяйства второй группы характеризуются в среднем худшими показателями потребления, чем домохозяйства первой группы.

Ковариационные матрицы для двух групп соответственно:

$$\hat{\Sigma}(1) = \begin{bmatrix} 0{,}08 & 0{,}03 & -0{,}03 \\ 0{,}03 & 0{,}009 & -0{,}006 \\ -0{,}03 & -0{,}006 & 0{,}05 \end{bmatrix} \text{ и } \hat{\Sigma}(2) = \begin{bmatrix} 1{,}16 & -0{,}02 & 0{,}22 \\ -0{,}02 & 0{,}08 & -0{,}02 \\ 0{,}22 & -0{,}02 & 0{,}07 \end{bmatrix}.$$

Суммарная ковариационная матрица имеет вид:

$$\hat{\Sigma} = \frac{1}{n_1 + n_2 - 2}\left(n_1\hat{\Sigma}(1) + n_2\hat{\Sigma}(2)\right) = \begin{bmatrix} 0{,}87 & 0{,}003 & 0{,}14 \\ 0{,}003 & 0{,}06 & -0{,}02 \\ 0{,}14 & -0{,}02 & 0{,}08 \end{bmatrix}.$$

Находим матрицу $\hat{\Sigma}^{-1}$, обратную к матрице $\hat{\Sigma}$:

$$\hat{\Sigma}^{-1} = \begin{bmatrix} 1{,}65 & -0{,}98 & -3{,}17 \\ -0{,}98 & 17{,}73 & 5{,}74 \\ -3{,}17 & 5{,}74 & 19{,}67 \end{bmatrix}.$$ Коэффициенты дискриминантной функции $\mathbf{w} = \hat{\Sigma}^{-1}(\hat{\mathbf{a}}_1 - \hat{\mathbf{a}}_2) = \begin{bmatrix} 2{,}76 \\ -5{,}28 \\ 6{,}97 \end{bmatrix}$.

Далее находим $\frac{1}{2}(\hat{\mathbf{a}}_1 + \hat{\mathbf{a}}_2) = \begin{bmatrix} 7{,}77 \\ 0{,}42 \\ 1{,}43 \end{bmatrix}$.

Классифицируем первое наблюдение:

$$\begin{bmatrix} 5{,}5 \\ 0{,}05 \\ 1{,}02 \end{bmatrix} - \begin{bmatrix} 7{,}77 \\ 0{,}42 \\ 1{,}43 \end{bmatrix} = \begin{bmatrix} -2{,}27 \\ -0{,}37 \\ -0{,}41 \end{bmatrix} \text{ и далее } \begin{bmatrix} -2{,}27 \\ -0{,}37 \\ -0{,}41 \end{bmatrix}^T \cdot \begin{bmatrix} 2{,}76 \\ -5{,}28 \\ 6{,}97 \end{bmatrix} = -7{,}16 < 0,$$

следовательно, наблюдение классифицируем во вторую группу неблагополучных.

Классифицируем второе наблюдение:

$$\begin{bmatrix} 10{,}0 \\ 0{,}32 \\ 2{,}62 \end{bmatrix} - \begin{bmatrix} 7{,}77 \\ 0{,}42 \\ 1{,}43 \end{bmatrix} = \begin{bmatrix} 2{,}23 \\ -0{,}10 \\ 1{,}19 \end{bmatrix} \text{ и далее } \begin{bmatrix} 2{,}23 \\ -0{,}10 \\ 1{,}19 \end{bmatrix}^T \cdot \begin{bmatrix} 2{,}76 \\ -5{,}28 \\ 6{,}97 \end{bmatrix} = 14{,}97 > 0$$

и домохозяйство классифицируется в первую группу — благополучных домохозяйств.

Пример 4.2. Имеются данные [14] по 43-м обанкротившимся фирмам, собранные за два года до наступления банкротства: x_1 — /наличные деньги/совокупный долг, x_2 — чистая прибыль/суммарные активы, x_3 — оборотные активы/краткосрочные обязательства, x_4 — оборотные активы/чистая выручка от продаж, x_5 — 0, если фирма обанкротилась и 1, если не обанкротилась.

Фирма принадлежит к одному из двух классов: банкрот или не банкрот.

Требуется классифицировать фирмы с характеристиками: а) x_1 =0,3; x_2 = 0,11; x_3 = 2; x_4 =0,14; а) x_1 =0,23; x_2 = 0,1; x_3 = 1,0; x_4 =0,4.

Решение.

Воспользуемся для вычислений пакетом *Statistica*. Проверим сначала наличие линейной зависимости переменных друг от друга. Вычислим корреляционную матрицу для переменных (Statistics\Basic Statistics\Correlation matrices):

Correlations (example 4_2)
Marked correlations are significant at p < ,05000
N=43 (Casewise deletion of missing data)

Variable	x_1	x_2	x_3	x_4
x_1	1,00	0,86	0,65	0,01
x_2	0,86	1,00	0,51	0,09
x_3	0,65	0,51	1,00	0,10
x_4	0,01	0,09	0,10	1,00

Поскольку обнаружена значимая линейная зависимость между переменными x_1 и x_2, а также x_1 и x_3, то исключим x_1 из списка дискриминантных переменных, так как она не дает новой информации, помимо той, которая содержится в переменных x_2 и x_3.

Оставшиеся три переменные проверим на нормальность распределения. Воспользуемся Graphs\Categorized Graphs\ Scatterplots, а также группировочной переменной x_5. Зададим опцию Overlaid и получим, например, для переменных x_2 и x_3 график вида:

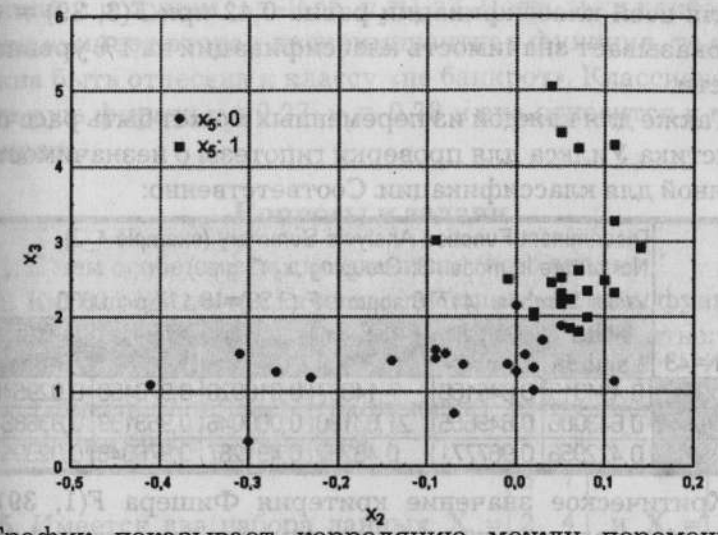


График показывает корреляцию между переменными внутри групп. В частности, предположение о двумерном нормальном распределении внутри каждой группы фирм, вероятно подтверждается для переменных x_2 и x_3.

Не является нормальным двумерное распределение для переменных x_3 и x_4. Однако это не фатально для проведения дискриминантного анализа.

Проверка предположения о равенстве ковариационных матриц выполняется с помощью дисперсионного анализа — опция Statistics\ANOVA\One-way ANOVA. М-тест Бокса отвергает гипотезу (χ^2=38,6) о равенстве ковариационных матриц двух групп на 5% уровне значимости, что связано с его чувствительностью к нормальности распределения переменных. Анализ по переменным показывает, что дисперсия неоднородна по группам для переменных x_2 и x_3. Тем не менее проведем дискриминацию по переменным x_2, x_3 и x_4.

Получим дискриминантные функции:

$y_1 = -13{,}52\,x_2 + 2{,}47\,x_3 + 12{,}16\,x_4 - 5{,}71;$

$y_2 = -1{,}70\,x_2 + 5{,}02\,x_3 + 10{,}34\,x_4 - 9{,}77.$

При этом апостериорные вероятности попадания в каждый из классов соответственно равны 0,47 и 0,53. Лямбда Уил-

кса для всей классификации равна 0,42 при $F(3, 39) = 18{,}12$, что показывает значимость классификации на 1% уровне значимости.

Также для каждой из переменных может быть рассчитана статистика Уилкса для проверки гипотезы о незначимости переменной для классификации. Соответственно:

Discriminant Function Analysis Summary (example 4_2)
No. of vars in model: 3; Grouping: x_5 (2 grps)
Wilks' Lambda: ,41778 approx. F (3,39)=18,117 p< ,0000

N=43	Wilks' Lambda	Partial Lambda	F-remove (1,39)	p-level	Toler.	1-Toler. (R-Sqr.)
x_2	0,494310	0,845186	7,14367	0,010926	0,970460	0,029540
x_3	0,643086	0,649655	21,03190	0,000046	0,963139	0,036861
x_4	0,422956	0,987771	0,48282	0,491267	0,976048	0,023952

Критическое значение критерия Фишера $F(1, 39)=4{,}07$ позволяет сделать вывод о незначимости для классификации только переменной x_4.

Таблица сопряженности «Факт — Прогноз» имеет вид:

Classification Matrix (example 4_2)
Rows: Observed classifications
Columns: Predicted classifications

Group	Percent Correct	G_1:0 p=,46512	G_2:1 p=,53488
G_1:0	90,0000	18	2
G_2:1	100,0000	0	23
Total	95,3488	18	25

В среднем процент правильных предсказаний равен 95,3%, следовательно, с точки зрения предсказания принадлежности к классам нарушение предположений о равенстве ковариационных матриц не несет большого вреда. Неправильно классифицированы во вторую группу два наблюдения из первой группы.

Классифицируем первую фирму. Для этого подставим значения ее характеристик в каждую из дискриминантных

функций и получим y_1=-0,55, y_2=1,53. Поскольку наибольшее значение имеет вторая дискриминантная функция, то фирма должна быть отнесена к классу «не банкрот». Классифицируем вторую фирму: y_1=0,27, y_2=-0,78 и она относится к классу «банкрот».

Вопросы и задачи

1. В чем особенности дискриминантного анализа?
2. Как определяется качество дискриминантных функций?
3. В чем суть непараметрического дискриминантного анализа?
4. Приведите пример (графически), когда дискриминантная функция будет нелинейной.
5. Имеется два набора данных $\mathbf{X}_1 = \begin{bmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{bmatrix}$ и $\mathbf{X}_2 = \begin{bmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{bmatrix}$, для которых $\bar{\mathbf{x}}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$, $\bar{\mathbf{x}}_2 = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$ и $\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$. Вычислите линейную дискриминантную функцию. Классифицируйте наблюдение $\mathbf{x}_0 = [2 \quad 7]$.
6. Деятельность двенадцати машиностроительных предприятий характеризуется показателями рентабельности (x_1, %) и производительности труда (x_2, тыс руб./чел.). Первые 4 предприятия имеют высокий уровень организации управления, а следующие 5 предприятий — низкий.

Требуется обосновать и реализовать подходящий метод классификации последних трех предприятий.

№ п.п.	Группы предприятий	x_1	x_2
1	Высокий	23,4	9,1
2		19,1	6,6
3		17,5	5,3
4		17,2	10,0

№ п.п.	Группы предприятий	x_1	x_2
5	Низкий	5,4	4,3
6		6,6	5,5
7		8,0	5,7
8		9,7	5,5
9		9,1	6,6
10	Подлежат классификации	9,9	7,4
11		14,2	9,4
12		12,9	6,7

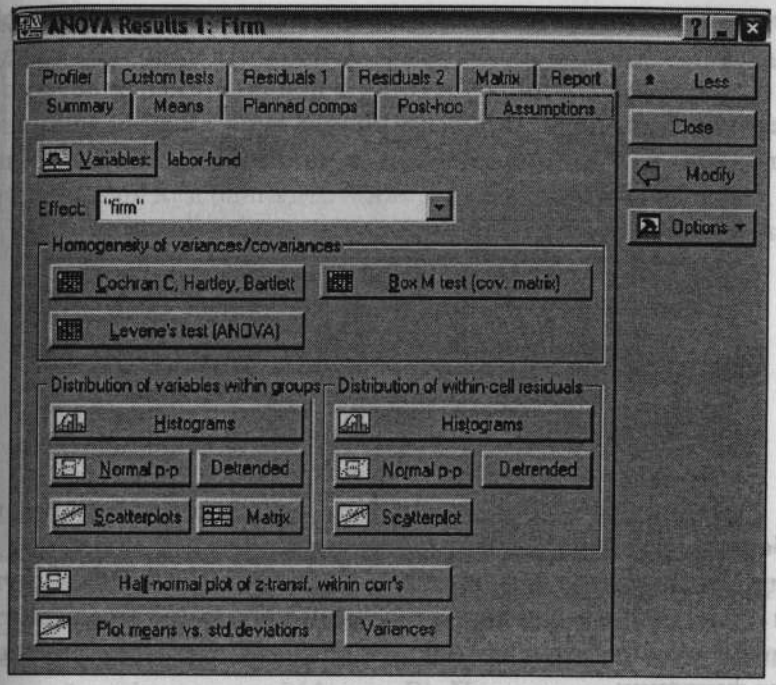
7. Как определяется количество дискриминантных функций?

8. Суть оптимального байесовского правила классификации.

Задание к лабораторному практикуму
Время выполнения — 4 часа

1. В файле firm.sta имеются данные по 12 предприятиям, характеризуемым тремя экономическими показателями: *labor* — производительность труда, *defect* — удельный вес потерь от брака (%) и *fund* — фондоотдача активной части основных производственных фондов. Из этих предприятий выделены две обучающие выборки (переменная *firm*), первая из которых включает 4 предприятия группы А, а вторая 5 предприятий группы В. Требуется классифицировать в одну из групп А или В оставшиеся три предприятия.

2. Перед выполнением дискриминантного анализа необходимо убедиться в том, что переменные, характеризующие предприятия, являются нормально распределенными и дисперсии и ковариации этих переменных внутри групп однородны. Для этого используется дисперсионный анализ. Необходимые опции реализованы в модуле ANOVA.



Выполнив опцию Statistics\ANOVA и выбрав One-way ANOVA, задайте лист зависимых (dependent) переменных *labor*, *defect*, *fund* и независимую переменную (factor) *firm*. Нажав ОК, выберите внизу появившегося окна опцию More results и затем вкладку Assumptions.

Затем, выбрав один из тестов в группе Homogeneity of variances/covariances (например, М-тест Бокса) путем нажатия соответствующей кнопки, получим результаты, которые убеждают нас в однородности дисперсий и ковариаций внутри двух групп.

Для проверки на нормальность распределения воспользуйтесь группой кнопок Distribution of variables within groups, например, графиками поля рассеяния: Scatterplots.

3. Выполните дискриминантный анализ имеющихся 9 предприятий, воспользовавшись меню Statistics\Multivariate Exploratory Techniques\Discriminant Analysis и указав в появив-

шемся окне в качестве группировочной переменной (Grouping variable) *firm*, а в качестве независимых (Independent variable list) остальные *labor*, *defect* и *fund*. В появившемся окне нажмите кнопку Summary. Получим результаты дискриминантного анализа по каждой переменной, в частности, лямбды Уилкса как для всей дискриминации, так и отдельно для каждой переменной и значимость переменных для классификации.

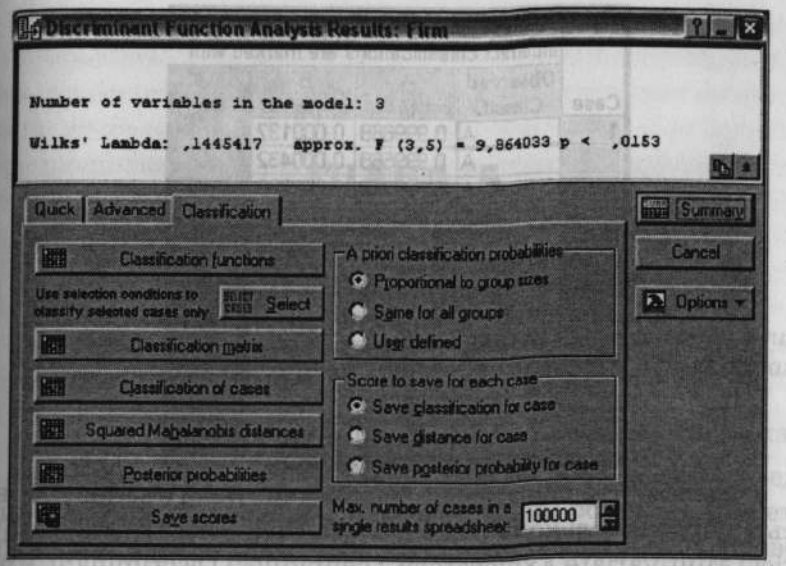
Discriminant Function Analysis Summary (Firm)
No. of vars in model: 3; Grouping: firm (2 grps)
Wilks' Lambda: ,14454 approx. F (3,5)=9,8640 p< ,0153

N=9	Wilks' Lambda	Partial Lambda	F-remove (1,5)	p-level	Toler.	1-Toler. (R-Sqr.)
labor	0,181529	0,796246	1,279471	0,309315	0,687015	0,312985
defect	0,153153	0,943771	0,297895	0,608664	0,877639	0,122361
fund	0,163325	0,884994	0,649754	0,456810	0,627656	0,372345

Вернувшись в окно Discriminant Function Analysis Results, выберите вкладку Classification и затем Classification functions. Получим значения коэффициентов дискриминантных функций с послеопытными вероятностями попадания предприятия в одну из групп. Должны получиться следующие дискриминантные функции:

$$f_A = -54{,}87 + 9{,}13labor + 6{,}39defect + 10{,}56fund;$$
$$f_B = -25{,}18 + 6{,}42labor + 11{,}07defect + 3{,}35fund.$$

Если в этом же окне выбрать опцию Classification matrix, получим матрицу по, строкам которой фактическая классификация, а по столбцам — полученная по модели. В идеальном случае они должны совпадать и матрица должна иметь диагональный вид при 100%-ных корректных наблюдениях 100%.



Далее, использовав опции Classification of cases или Posterior probabilities, получим соответственно классификацию по наблюдениям и вероятности отнесения каждого наблюдения к каждой из двух групп (A или B). Причем классифицированы будут и последние 3 наблюдения, для которых мы не имели первоначально информации о том, к какой из групп они относятся (наблюдения 10 и 11 — к группе B, а 12 — к группе A). Также можно получить квадрат расстояния Махаланобиса от центра каждой из групп с помощью опции Squared Mahalanobis distances.

Case	Posterior Probabilities (Firm) Incorrect classifications are marked with * Observed Classif.	A p=,44444	B p=,55556
1	A	0,999868	0,000132
2	A	0,999568	0,000432
3	A	0,999740	0,000260
4	A	0,999989	0,000011
5	B	0,000478	0,999522
6	B	0,000000	1,000000
7	B	0,009825	0,990175
8	B	0,010251	0,989749
9	B	0,000000	1,000000
10	---	0,000492	0,999508
11	---	0,000935	0,999065
12	---	1,000000	0,000000

4. Выполните пошаговый дискриминатный анализ. Вернитесь к первоначальному окну дискриминантного анализа (Statistics\Multivariate Exploratory Techniques\Discriminant Analysis) и поставьте галочку напротив опции Advanced options. Нажмите ОК. Выберите вкладку Advanced (обратите внимание на возможности изменения значений F критерия для включения/исключения переменной и вида отображения — конечного результата или результатов по шагам) и метод (Method) пошагового анализа: Forward (включения) или Backward (исключения). При этом опция Standard относится к стандартному алгоритму анализа, выполненному нами в предыдущем пункте. Нажмите ОК и получите окно результатов, имеющее такой же, как и в п. 3, вид.

Выполните пошаговый анализ методом последовательного включения переменных и методом исключения.

5. Получите результаты в п. 2–4. Сделайте содержательные выводы по результатам всех выполненных расчетов.

ГЛАВА 5
Классификация без обучения. Кластерный анализ

Не в совокупности ищи единства, но более — в единообразии разделения.

К. Прутков. Мысли и афоризмы

Методы кластерного анализа позволяют решать следующие задачи: проведение классификации объектов с учетом признаков, отражающих сущность, природу объектов; проверка выдвигаемых предположений о наличии некоторой структуры в изучаемой совокупности объектов; построение новых классификаций для слабоизученных явлений, когда необходимо установить наличие связей внутри совокупности.

Общая постановка задачи классификации совокупности объектов $O_1, O_2, ..., O_n$ в условиях отсутствия обучающих выборок состоит в требовании разбиения этой совокупности на некоторое число (заранее известное или нет) однородных в определенном смысле классов. При этом исходная информация о классифицируемых объектах представлена либо значениями многомерного признака (по каждому объекту в отдельности), либо матрицей попарных расстояний между объектами, а понятие однородности основано на предположении, что геометрическая близость двух или нескольких объектов означает их сходство.

В зависимости от наличия и характера априорной информации о природе искомых классов и от конечных целей многомерной классификации при отсутствии обучающих выборок используют один из трех подходов: 1) методы расщепления смесей вероятностных распределений (каждый класс интер-

претируется как параметрически заданная одномодальная генеральная совокупность при неизвестном значении определяющего ее параметра, а классифицируемые наблюдения — как выборка из смеси таких генеральных совокупностей), 2) методы кластерного анализа, 3) классификационные процедуры иерархического типа (главная цель — получение наглядного представления о стратификационной структуре всей классифицируемой совокупности).

Как отмечается в [11], большинство методов кластерного анализа являются эвристическими и представляют собой довольно простые процедуры, что позволяет свести к минимуму ошибки при трактовке результатов анализа. Однако *необходимо иметь в виду*, что кластерные методы размещают объекты по группам, которые могут существенно различаться по составу при использовании различных методов кластеризации. Кластерный метод привносит структуру в данные (хотя цель кластеризации в нахождении структуры), и эта структура может не совпадать с искомой, «реальной». Ключевым является умение отличать «реальные» группировки от навязанных методами кластерного анализа.

5.1. Параметрический случай классификации без обучения. Расщепление смесей вероятностных распределений

Рассмотрим пример смеси распределений из [2. С. 182].

В отдел технического контроля (ОТК) поступают партии изделий, составленные с помощью случайного извлечения изделия из объединенной продукции, произведенной на разных станках А и В. Изделия контролируются по некоторому количественному параметру ξ мм, так что результатом контроля i-го изделия партии является число x_i мм (изделия на станках не маркируются и в ОТК не известно, на каком именно станке произведено каждое из них). Производительность станка А в

1,5 раза больше производительности станка В. Задано номинальное значение контролируемого параметра a=65 мм и известно, что точность работы станков характеризуется одинаковой величиной среднеквадратических отклонений $\Sigma_A=\Sigma_B=1$ мм^2. Предположим, что потом выяснилось, что станок А был настроен правильно и производил изделие с номинальным значением параметра $E\xi_A$= 65 мм, а настройка станка В была сбита $E\xi_B$= =67 мм. Известно также, что распределение размеров изделий, произведенных на каком-то определенном станке, описывается нормальным законом.

Очевидно, анализируемая в ОТК по наблюдениям x_1, x_2, ... генеральная совокупность будет состоять из смеси двух нормальных генеральных совокупностей, одна из которых представляет продукцию станка А и описывается плотностью

$f_A(x)=\frac{1}{\sqrt{2\pi}\sigma_A}e^{\frac{(x-a_A)^2}{2\sigma_A^2}}$, а другая — продукцию станка В и описывается плотностью $f_B(x)=\frac{1}{\sqrt{2\pi}\sigma_B}e^{\frac{(x-a_B)^2}{2\sigma_B^2}}$.

Обозначим $\theta_j=(a_j,\sigma_j^2)$, где $j=A, B$, а удельный вес изделий станка j через p_j. Тогда уравнение функции плотности, описывающей закон распределения анализируемого признака ξ во всей (объединенной) генеральной совокупности, в виде:

$$f(x)=p_Af_A(x,\theta_A)+p_Bf_B(x,\theta_B).$$

Так как в объединенной генеральной совокупности продукции станка А в 1,5 раза больше, чем продукции станка В, а также $\Sigma_A=\Sigma_B=1$ мм^2, a_A= 65 мм, a_B= 67 мм, то имеем:

$$f(x)=0{,}6\frac{1}{\sqrt{2\pi}}e^{\frac{(x-65)^2}{2}}+0{,}4\frac{1}{\sqrt{2\pi}}e^{\frac{(x-67)^2}{2}}.$$

Получили так называемую смесь вероятностных распределений (рис. 5.1).

Если сотрудники ОТК или потребители изделий захотят по наблюдениям x_1, x_2, ... определить, на каком именно станке произведено каждое из них, то возникает одна из задач классификации наблюдений в условиях отсутствия обучающих выборок.

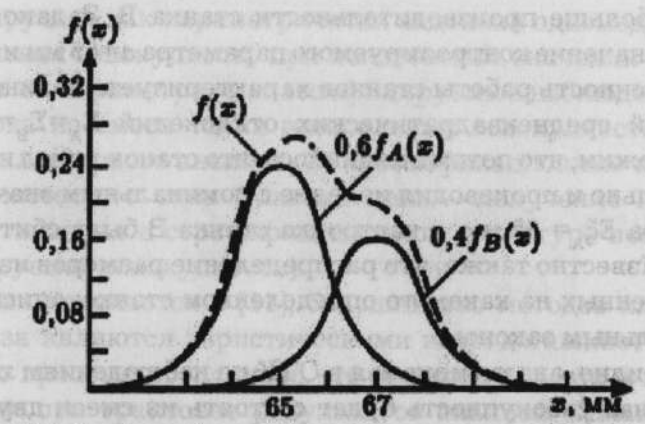


Рис. 5.1. *Графики функций плотности и смеси распределений*

Рассмотрим далее формализацию задачи расщепления смеси распределений.

Пусть имеется двухпараметрическое семейство p–мерных плотностей распределения: $F = \{f_w(X;\theta(W))\}$,

где w — параметр, определяющий специфику общего вида каждого компонента распределения в смеси, $\theta(w)$ — параметр (в общем случае — вектор).

Пусть $\Psi = \{\psi(W)\}$ — семейство смешивающих функций распределения.

Если смесь состоит из конечного числа k компонентов f_w, то априорные вероятности появления элементов класса w: $\psi(w) = \pi_w$, $w=1,...,k$.

Функция $f(X)$ называется *смесью вероятностных распределений*, если она представима в виде $f(X) = \int f_w(X;\theta(w))\psi(w)dw$.

В случае конечного числа классов k и однотипности компонентов распределений $f_j(X;\theta_j)$, т.е. принадлежности одному общему семейству $f_j(X;\theta_j) \in \{\tilde{f}(X;\theta)\}$, модель смеси будет иметь вид

$$f(X) = \sum_{j=1}^{k} \pi_j \tilde{f}_j(X,\theta(j)). \qquad (5.1)$$

Пусть имеется выборка классифицируемых наблюдений $X_1,...,X_n$, извлеченная из генеральной совокупности, являющейся смесью вида (5.1). Необходимо построить статистические оценки: для числа компонентов смеси k, их удельных весов $\pi_1,...,\pi_k$ для каждого из компонентов $f_w(X;\theta(w))$ анализируемой смеси.

Часто число компонентов смеси k известно, известны $\pi_1,...,\pi_k$, тогда нужно только построить оценки для $f_w(X;\theta(w))$.

В схеме *автоматической классификации*, опирающейся на *модель смеси распределений*, заданы классы функций $f_1(X;\theta_1), f_2(X;\theta_2),...$ Неизвестные значения параметров θ_1, θ_2, ... и параметров k, $\pi_1,...,\pi_k$ могут быть оценены не по обучающим выборкам (их у нас нет), а по классифицируемым наблюдениям $X_1,...,X_n$ с помощью одного из известных методов статистического оценивания параметров (метод максимального правдоподобия, метод моментов и т.п.).

После этого имеем схему дискриминантного анализа, и процесс классификации производится так же, как и в параметрическом дискриминантном анализе: относим наблюдение X_0 к классу j^*, если $\hat{\pi}_{j^*} f_{j^*}(X_0;\hat{\theta}_{j^*}) = \max\limits_{1\le j\le k} \{\hat{\pi}_j f_j(X_0,\hat{\theta}_j)\}$.

На практике получить оценки $\hat{\theta}_1,\hat{\theta}_2,...,\hat{k},\hat{\pi}_1,...,\hat{\pi}_k$ весьма сложно.

5.2. Непараметрический случай классификации без обучения: кластерный анализ

Постановка задачи автоматической классификации.

Пусть мы не располагаем обучающими выборками и также отсутствует информация о характере распределения наблюдений $X_i, i=1,...,n$ внутри каждого из классов.

В общей постановке проблема автоматической классификации объектов заключается в том, чтобы всю совокупность объектов, статистически представленную в виде матриц (1.1) или (1.2), разбить на сравнительно небольшое число (заранее известное или нет) однородных, в определенном смысле, групп или классов.

Исходные данные — это точки $X_i, i=1,...,n$ в p-мерном признаковом пространстве. Выбор переменных кластеризации (признаков) является одним из наиболее важных элементов анализа. Идеально, если в основе выбора лежит экономическая теория. Однако на практике часто теория, обосновывающая классификацию, не сформулирована или спорна. В этом случае не следует руководствоваться принципом, «чем больше, тем объективнее» и включать в анализ как можно большее количество переменных. Это может привести к коррелированности переменных и, по сути, к взвешиванию этих переменных. Часто одновременно с кластерным анализом используются метод главных компонент и факторный анализ для уменьшения размерности данных $\left(x^{(1)}, x^{(2)}, ..., x^{(p)}\right)$ и отбора самых информативных (гл. 6, 7).

Геометрическая близость точек означает, что они находятся на сравнительно небольших расстояниях друг от друга. Полученные в результате разбиения классы называют кластерами (cluster), т.е. *кластер* — это группа элементов, обладающих каким-то общим свойством, а методы нахождения кластеров называются кластерным анализом. Впервые эти термины введены Трайоном (*R. Tryon*, 1939 г.) [1].

При решении задачи классификации важно, как именно действует исследователь. Возможно группирование наблюдений в интервалы, в результате исследуемая совокупность объектов разбивается на некоторое число групп. Или же выявляются области повышенной плотности наблюдений. В первой постановке задача всегда имеет решение, во втором случае может оказаться, что множество исходных наблюдений не обнаруживает естественного расслоения на кластеры (например, образует общий кластер).

Расстояния между отдельными объектами и меры близости объектов друг к другу.

Наиболее трудным и наименее формализованным в задаче автоматической классификации является понятие *однородности объектов.*

Однородность определяется заданием правила вычисления либо расстояния d между объектами, либо заданием некоторой меры степени близости (сходства) ϖ тех же объектов. Ясно, что при этом необходимо сопоставление с некоторым пороговым значением, определяемым для каждого конкретного случая.

Также очевидно, что расстояние между объектами должно быть:

- симметрично $d(X_j, X_i) = d(X_i, X_j)$ и $\varpi(X_j, X_i) = \varpi(X_i, X_j)$,
- объект должен быть максимально похож сам на себя $\varpi(X_j, X_j) = \max_{1 \le i \le n} \varpi(X_i, X_j)$,
- степень сходства ϖ должна монотонно убывать с расстоянием $d(X_k, X_l) \ge d(X_i, X_j) \Rightarrow \varpi(X_k, X_l) \le \varpi(X_i, X_j)$.

От выбора меры сходства объектов часто зависит успешность кластерного анализа, поэтому этот выбор должен быть органической частью плана исследования, определяющегося теоретическим и практическим содержанием задачи классификации.

Виды расстояний между объектами.

Выбор расстояния является узловым моментом классификации, от которого решающим образом зависит разбиение объектов на классы при заданном алгоритме разбиения. В каждой конкретной задаче этот выбор определяется на основе статистической и экономической природы наблюдений X и полноты априорных сведений о характере их вероятностного распределения. Если, например, известно, что наблюдения извлекаются из нормально распределенных генеральных совокупностей с одной и той же матрицей ковариаций, то используется расстояние Махаланобиса.

На практике используются следующие виды расстояний между объектами.

- Расстояние Махаланобиса. Если компоненты $\left(x^{(1)}, x^{(2)}, ..., x^{(p)}\right)$ вектора наблюдений X зависимы и их значимость в решении вопроса о классификации объекта различна, то пользуются расстоянием типа $d_0(X_i, X_j) = \sqrt{\left((X_i - X_j)^T \Delta^T \Sigma^{-1} \Delta (X_i - X_j)\right)}$, где Σ — ковариационная матрица генеральной совокупности,

из которой извлекаются наблюдения X_i; Δ — некоторая симметричная неотрицательно определенная матрица весовых коэффициентов, которая чаще всего выбирается диагональной.

Другие три расстояния являются частными случаями расстояния Махаланобиса.

■ Евклидово расстояние. $d_E(X_i,X_j)=\sqrt{\sum_{l=1}^{p}(x_i^{(l)}-x_j^{(l)})^2}$.

Оно используется, если наблюдения извлекаются из генеральной совокупности, которая описывается многомерным нормальным законом распределения, причем X должны быть взаимно независимыми и иметь одинаковую дисперсию. Компоненты этих наблюдений $\left(x^{(1)},x^{(2)},...,x^{(p)}\right)$ должны быть однородными, т.е. одинаково важными для классификации. Признаковое пространство p должно совпадать с геометрическим (p не может быть больше 3, т.е. принимать значение 1, 2 или 3).

■ Взвешенное евклидовое расстояние.

$$d_{BE}(X_i,X_j)=\sqrt{\sum_{l=1}^{p}w_l(x_i^{(l)}-x_j^{(l)})^2}.$$

Компоненты $\left(x^{(1)},x^{(2)},...,x^{(p)}\right)$ неоднородны и степень их важности задают веса w_l, причем $0<w_l<1$, l=1,..,p.

Определение весов требует дополнительного исследования (экспертные опросы, специальные модели и др.), определение их по исходным данным не дает желаемого эффекта.

■ Расстояние Минковского $d_M(X_i,X_j)=\left(\sum_{l=1}^{p}\left|x_i^{(l)}-x_j^{(l)}\right|^g\right)^{1/g}$ и частные его случаи: при g=1 Хеммингово расстояние (манхеттенское или расстояние городских кварталов city-block)

$d_H(X_i,X_j)=\sum_{l=1}^{p}\left|x_i^{(l)}-x_j^{(l)}\right|$, которое используется в случае объектов, характеризуемых дихотомическими признаками.

При g=2 получаем евклидово расстояние.

В некоторых задачах используются содержательные числовые параметры, характеризующие взаимоотношения между объектами, например, матрица межотраслевого баланса для отраслей.

Отметим, что, несмотря на распространенность указанных видов расстояний, они *имеют существенный недостаток* — оценка сходства сильно зависит от различий в сдвигах (среднее значение для объекта по всем признакам — переменным) данных. Переменные (признаки), у которых одновременно велики абсолютные значения и стандартные отклонения, могут подавить влияние переменных с меньшими абсолютными размерами и стандартными отклонениями. Расстояния в метрических шкалах изменяются под воздействием преобразований шкалы измерения переменных, при которых не сохраняется ранжирование по евклидову расстоянию. Чтобы уменьшить различия в сдвигах, переменные стандартизируют вычитанием среднего и делением на среднеквадратическое отклонение.

Если необходимо установить сходство между объектами, описываемыми бинарными переменными, применяют различные коэффициенты ассоциативности [11, 14]. Их применение рассмотрено в примере 5.4.

Расстояние между классами объектов.

В некоторых методах кластерного анализа важно расстояние между группами объектов в целом. Пусть S_i — i-я группа, i=1,...,k, n_i — число объектов в этой группе, вектор $\bar{X}(i)$ — среднее арифметическое векторных наблюдений группы S_i («центр тяжести» группы), $\rho(S_l,S_m)$ — расстояние между группами S_l и S_m.

Некоторые виды расстояний (рис. 5.2):

■ Расстояние по принципу «ближнего соседа»

$$\rho_{\min}(S_l,S_m)=\min_{X_i\in S_l,X_j\in S_m}d(X_i,X_j).$$

Здесь d — расстояние между объектами.

■ Расстояние по принципу «дальнего соседа»

$$\rho_{\max}(S_l,S_m)=\max_{X_i\in S_l,X_j\in S_m}d(X_i,X_j).$$

■ Расстояние по «центрам тяжести» групп

$$\rho(S_l,S_m)=d(\bar{X}(l),\bar{X}(m)).$$

■ Расстояние по принципу «средней связи»

$$\rho_{cp}(S_l,S_m)=\frac{1}{n_l n_m}\sum_{X_i\in S_l}\sum_{X_j\in S_m}d(X_i,X_j).$$

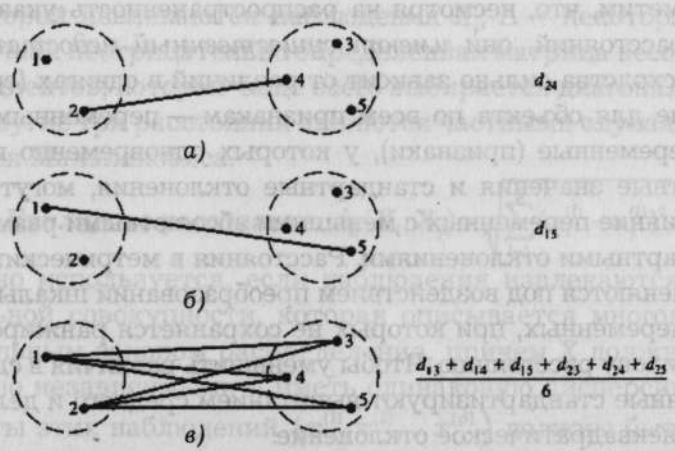


Рис. 5.2. *Расстояния между классами объектов: а) «ближний сосед», б) «дальний сосед», в) средней связи*

Обобщением всех приведенных выше расстояний является расстояние по Колмогорову:

$$\rho_\tau^{(k)}(S_l,S_m)=\left[\frac{1}{n_l n_m}\sum_{X_i\in S_l}\sum_{X_j\in S_m} d^\tau(X_i,X_j)\right]^{1/\tau}.$$

Если $\tau\to\infty$, то $\rho_\tau^{(k)}=\rho_{\max}$, если $\tau\to-\infty$, то $\rho_\tau^{(k)}=\rho_{\min}$, $\tau=1$, $\rho_\tau^{(k)}=\rho_{ср}$.

Для измерения расстояний между нормальными классами с номерами l и m используют так называемое информационное расстояние Каллбэка (Kullback)

$$\rho^2(S_l,S_m)=\frac{1}{2}\big(a(l)-a(m)\big)^T\big(\Sigma^{-1}(l)+\Sigma^{-1}(m)\big)\big(a(l)-a(m)\big)+$$

$$+\frac{1}{2}tr\Big\{\big(\Sigma(l)-\Sigma(m)\big)\big(\Sigma^{-1}(l)-\Sigma^{-1}(m)\big)\Big\},\ где\ a(l)=\bar{X}(l), a(m)=\bar{X}(m)\ .$$

В последней формуле Σ — ковариационная матрица наблюдений, tr — след матрицы.

Если анализируемые классы различаются только средними значениями, а ковариационные матрицы у них одинаковые $\Sigma(l)=\Sigma(m)=\Sigma$, то используется расстояние Махаланобиса:

$$\rho^2(S_l,S_m)=(a(l)-a(m))^T\Sigma^{-1}(a(l)-a(m)).$$

На практике две последние формулы используются и в тех случаях, когда распределение наблюдений внутри классов отличается от нормального с заменой средних и ковариационных матриц $a(j)$ и $\Sigma(j)$ их оценками $\hat{a}(j)$ и $\hat{\Sigma}(j)$, построенными по наблюдениям.

Оценка качества разбиения на классы.

Функционал качества разбиения определяется на множестве всех возможных классификаций, и наилучшей классификацией является та, на которой он достигает экстремума. Выбор меры качества разбиения осуществляется эмпирически. Рассмотрим некоторые наиболее распространенные функционалы.

Функционалы качества разбиения при заданном числе классов.

Пусть, как и ранее d — расстояние, S — некоторое разбиение наблюдений $X_i, i=1,\dots,n$ на заданное число классов k.

Меры (рис. 5.3):

- сумма внутриклассовых дисперсий

$$Q_1(S)=\sum_{l=1}^{k}\sum_{x_i\in S_l} d^2(X_i,\bar{X}(l)).$$

Стремятся минимизировать Q_1;

- сумма попарных внутриклассовых расстояний между элементами

$$Q_2(S)=\sum_{l=1}^{k}\sum_{X_i,X_j\in S_l} d^2(X_i,X_j).$$

Стремятся минимизировать Q_2, т.е. получить кластеры большей «плотности»;

- обобщенная внутриклассовая дисперсия характеризует степень рассеивания многомерных наблюдений одного класса около своего «центра тяжести»

$$Q_3(S)=\det\left(\sum_{l=1}^{k} n_l\hat{\Sigma}(l)\right)\ \text{или}\ Q_4=\prod_{l=1}^{k}\det\big(\hat{\Sigma}(l)\big)^{n_l}.$$

Функционалы Q_3 и Q_4 обычно используются, когда необходимо выяснить, можно ли сократить размерность пространства p.

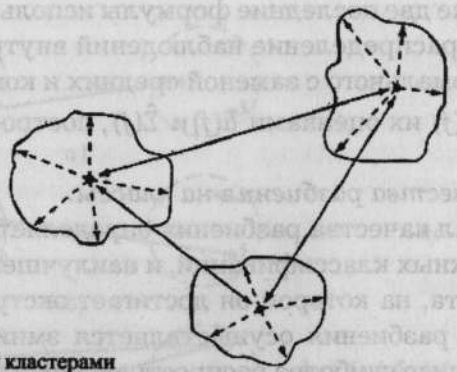


Рис. 5.3. *Кластерная диаграмма, показывающая между- и внутрикластерные вариации*

Функционалы качества разбиения при неизвестном числе классов.

В этом случае $Q(S)$ выбирается в виде некоторой комбинации из двух функционалов: $Z_\tau(S)$ — мера концентрации, $I_\tau(S)$ — средняя мера внутриклассового рассеивания. Такой подход называется подходом Колмогорова. Расчетные формулы

$$Z_\tau(S)=\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{n(x_i)}{n}\right)^\tau\right]^{1/\tau},$$ где $n(X_i)$ — число элементов в кластере, содержащем точку X_i; τ — параметр:

$$\tau=-1,\quad Z_{-1}(S)=\frac{1}{k};$$

$$\tau=0,\quad Z_0(S)=\sum_{i=1}^{k}\frac{n_i}{n}\log\frac{n_i}{n};$$

$$\tau=\infty,\quad Z_\infty(S)=\max_{1\le i\le k}\left(\frac{n_i}{n}\right);$$

$$\tau=-\infty,\quad Z_{-\infty}(S)=\min_{1\le i\le k}\left(\frac{n_i}{n}\right);$$

$$\tau=1,\quad Z_1(S)=\frac{1}{n^2}\sum_{i=1}^{k}n_i^2.$$

При любом τ мера $Z_\tau(S)$ имеет минимальное значение $Z_\tau(S)=\frac{1}{n}$ при разбиении множества на n классов и максимальное значение, равное единице, при объединении всех наблюдений в один кластер. Далее $I_\tau(S)=\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n(X_i)}\sum_{X_l\in S(X_i)}d^\tau(X_i,X_l)\right]^{1/\tau}$.

Можно использовать следующие линейные комбинации, например, $Q(S)=\alpha I_\tau(S)+\beta/Z_\tau(S)$, где α и β — параметры, например, $\alpha=\beta=1$ или $Q(S)=\left[I_\tau(S)\right]^\alpha\left[\frac{1}{Z_\tau(S)}\right]^\beta$.

Далее рассмотрим основные типы задач, решаемых с помощью кластерного анализа, и алгоритмы решения.

5.3. Основные типы задач кластер-анализа и основные типы кластер-процедур

В зависимости от количества наблюдений n, выделяют два типа задач кластер-анализа, условно $Б_1$ и $Б_2$ [1].

К типу $Б_1$ относят те задачи классификации, в которых число наблюдений несколько десятков. К типу $Б_2$ — задачи классификации, где число наблюдений порядка нескольких сотен, тысяч.

С точки зрения априорной информации об окончательном числе классов выделяют три класса задач:

(а) — число классов известно априори,

(б) — число классов неизвестно и подлежит оценке,

(в) — число классов неизвестно, но его определение не входит в условие задачи. Требуется построить так называемое иерархическое дерево исследуемой совокупности, т.е. дендрограмму.

В соответствии с такой классификацией задач кластерного анализа выделяют три основных типа кластер-процедур (рис. 5.4).

1. *Иерархические* процедуры, которые делятся на *агломеративные* и *дивизимные*. Предназначены для решения задач типа в). Формально могут применяться для задач и $Б_1$ и $Б_2$, но на практике реализуются конструктивно для задач типа $Б_1$.

2. *Параллельные* процедуры, предназначены для решения задач типов $Б_1$(а) и $Б_1$(б) и реализуются с помощью итерационных алгоритмов, на каждом шаге которых одновременно используются все имеющиеся наблюдения.

3. Процедуры *последовательные*. Предназначены для решения задач $Б_2$(а) и $Б_2$(б). Реализуются в виде итерационных алгоритмов, на каждом шаге которых используется лишь небольшая часть исходных наблюдений, а также результат разбиения на предыдущем шаге.

5.4. Иерархические процедуры

Принцип работы агломеративных (дивизимных) процедур состоит в последовательном объединении (разделении) групп элементов сначала самых близких (далеких), а затем все более отдаленных друг от друга (приближенных друг к другу). При этом агломеративные процедуры начинают обычно с объединения отдельных элементов, а дивизимные — с разъединения всей исходной совокупности наблюдений.

Иерархические процедуры дают более полный и тонкий анализ исследуемого множества наблюдений: структуру исследуемого множества наблюдений, наглядную интерпретацию результатов. К недостаткам иерархических процедур следует отнести: громоздкость их вычислительной реализации, их применение при числе наблюдений несколько сотен либо невозможно, либо нецелесообразно. Кроме того, имеется достаточно большое количество примеров, когда результаты разбиения весьма далеки от оптимальных.

Некоторые иерархические процедуры.

Агломеративный иерархический алгоритм «ближайшего соседа» (или «одиночной связи»). Алгоритм использует расстояние между кластерами по правилу ближайшего соседа. На первом шаге алгоритма каждое наблюдение рассматривается как отдельный кластер. Далее происходит объединение двух самых близких кластеров и соответственно пересчитывается матрица расстояний, размерность которой снижается на единицу. Работа алгоритма заканчивается, когда все наблюдения объединены в один класс. Понятно, что два элемента попадают в один кластер, если существует соединяющая их цепочка близких между собой элементов. Метод приводит к появлению «цепочек», т.е. к образованию больших продолговатых кластеров.

Агломеративные иерархические алгоритмы «средней связи» и «полной связи» (или «дальнего соседа»). Отличаются от предыдущего алгоритма способами вычисления расстояния между классами. В «среднем» берется среднее расстояний, а в «полной связи» — по принципу «дальнего соседа».

К — обобщенная иерархическая процедура (по Колмогорову). Используется обобщенное расстояние Колмогорова между классами и поэтому все иерархические процедуры можно рассматривать как частные случаи этой процедуры.

Иерархические процедуры, использующие понятие порога. Отличие этих процедур от остальных — дополнительное задание последовательности порогов $c_1, c_2, ..., c_t$. На первом шаге алгоритма попарно объединяются элементы, расстояние между которыми не превосходит величины c_1. На втором шаге объединяются группы элементов, расстояние между которыми не превосходит c_2 и т.д. Недостатком такой процедуры является сложность обоснованного выбора порогов.

Метод Уорда. На первом шаге алгоритма метода каждый кластер полагается состоящим из одного объекта. Далее объединяются два ближайших кластера. Для них определяются средние значения каждого признака и рассчитывается сумма квадратов отклонений $V_k = \sum_{i=1}^{n_k} \sum_{j=1}^{p} (x_{ij} - \overline{x}_{jk})^2$, где k — номер кластера; i — номер объекта; j — номер признака; p — количество признаков, характеризующих каждый объект; n_k — количество

объектов в k-м кластере. В дальнейшем на каждом шаге работы алгоритма объединяются те объекты или кластеры, которые дают наименьшее приращение величины V_k. Метод приводит к образованию кластеров приблизительно равных размеров с минимальной внутрикластерной вариацией. В итоге все объекты оказываются объединенными в один кластер.

Таким образом, все иерархические агломеративные методы последовательно объединяют наиболее схожие объекты и требуют $n-1$ шагов. Полученную последовательность объединений кластеров можно представить визуально в виде древовидной диаграммы, называемой дендрограммой. Недостатками этих методов кластерного анализа является то, что объекты распределятся по кластерам за один проход и поэтому плохое начальное разбиение данных не может быть изменено на последующих шагах кластеризации. Кроме того, они могут (за исключением метода одиночной связи) давать разные решения в результате простого переупорядочивания объектов, что не позволяет считать группировки устойчивыми.

5.5. Последовательные кластер-процедуры

В случаях, когда число n классифицируемых наблюдений велико, иерархические процедуры крайне трудоемки и надо использовать алгоритмы, на каждом шаге которых обсчитывается лишь небольшая часть исходных наблюдений, например одно из них.

Метод k-средних принадлежит к группе итеративных методов эталонного типа (*Дж. Мак-Куин*, 1967).

Пусть имеется n наблюдений, характеризующихся p признаками $x_{i1}, x_{i2}, \ldots, x_{ip}$, $i = 1, \ldots, n$. Эти наблюдения необходимо разбить на заданное число классов $k \le n$.

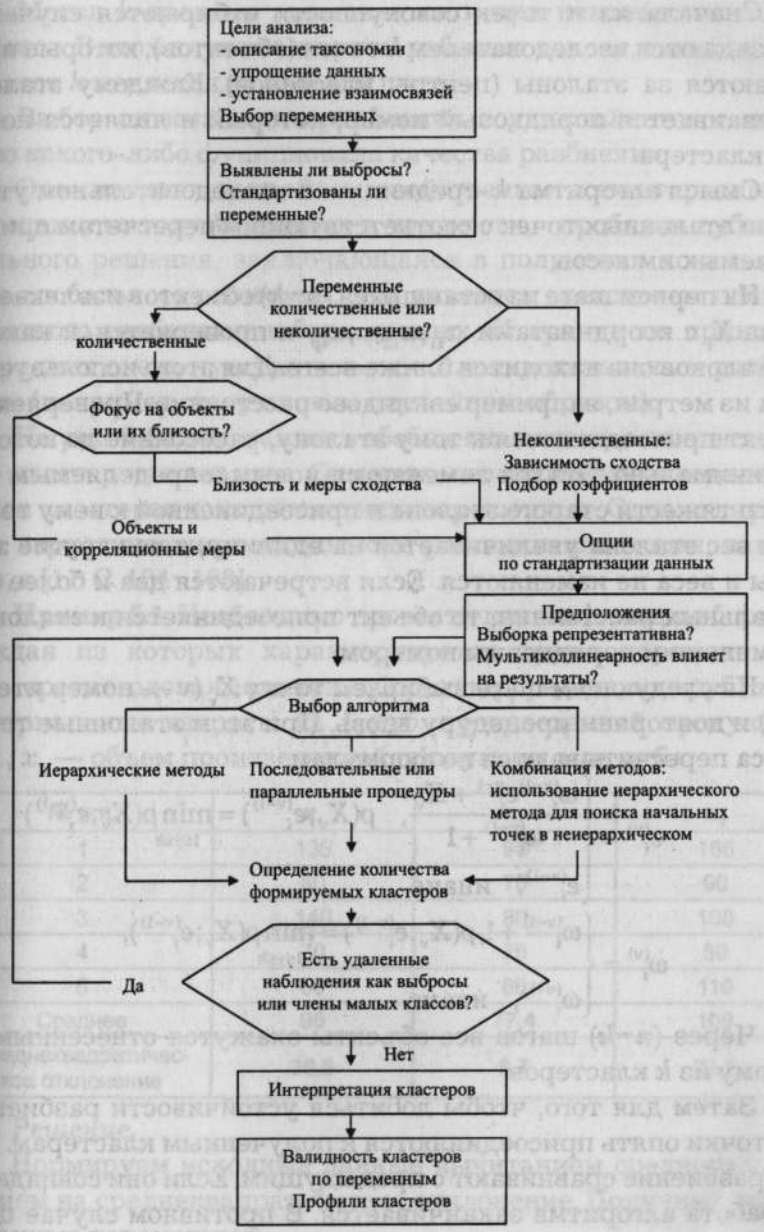


Рис. 5.4. *Блок-схема принятия решений в кластерном анализе*

Сначала из n точек совокупности отбираются случайно или задаются исследователем k точек (объектов), которые принимаются за эталоны (центры кластеров). Каждому эталону присваивается порядковый номер, который и является номером кластера.

Смысл алгоритма k-средних — в последовательном уточнении эталонных точек с соответствующим пересчетом приписываемых им весов.

На первом шаге из оставшихся $(n-k)$ объектов извлекается точка X_i с координатами $x_{i1}, x_{i2}, ..., x_{ip}$ и проверяется, к какому из эталонов она находится ближе всего. Для этого используется одна из метрик, например евклидово расстояние. Проверяемый объект присоединяется к тому эталону, расстояние до которого минимально. Эталон заменяется новым, определяемым как центр тяжести старого эталона и присоединенной к нему точки X_i, и вес эталона увеличивается на единицу, а все другие эталоны и веса не изменяются. Если встречаются два и более минимальных расстояния, то объект присоединяется к эталону с наименьшим порядковым номером.

На следующем шаге выбираем точку X_v (v — номер итерации) и повторяем процедуру вновь. При этом эталонные точки и веса пересчитываются по формулам

$$e_i^{(v)} = \begin{cases} \dfrac{\omega_i^{(v-1)} e_i^{v-1} + X_v}{\omega_i^{(v-1)} + 1}, & \rho(X_v, e_i^{(v-1)}) = \min\limits_{1 \le j \le k} \rho(X_v; e_j^{(v-1)}) \\ e_i^{(v-1)}, & \text{иначе} \end{cases}$$

$$\omega_i^{(v)} = \begin{cases} \omega_i^{(v-1)} + 1, \rho(X_v, e_i^{(v-1)}) = \min\limits_{1 \le j \le k} \rho(X_v; e_j^{(v-1)}), \\ \omega_i^{(v-1)}, & \text{иначе} \end{cases}$$

Через $(n-k)$ шагов все объекты окажутся отнесенными к одному из k кластеров.

Затем для того, чтобы добиться устойчивости разбиения, все точки опять присоединяются к полученным кластерам. Новое разбиение сравнивают с предыдущим. Если они совпадают, то работа алгоритма заканчивается. В противном случае цикл повторяется.

Метод k-средних применяется и при неизвестном числе классов. В этом случае алгоритм работает так же, только число классов k меняется на каждом шаге.

Выбор оптимального разбиения осуществляется с помощью какого-либо функционала качества разбиения.

Одна из главных проблем, свойственная всем итеративным методам и методу k-средних в частности,– проблема субоптимального решения, заключающаяся в получении локального, а не глобального оптимума, что является следствием плохого исходного разбиения набора данных. Итерации по принципу k-средних очень чувствительны к плохим начальным разбиениям (например, полученным случайным образом).

На рис. 5.4 представлена обобщенная схема основных этапов кластерного анализа. Отметим, что проблема определения числа кластеров находится среди нерешенных. Однако некоторые эвристические подходы к ее решению изложены, например, в [11. С. 184–188].

Пример 5.1. Необходимо провести классификацию 5 фирм, каждая из которых характеризуется тремя переменными: x_1 — среднегодовая величина оборотных средств, млн руб., x_2 — материальные затраты на 1 руб. произведенной продукции, коп., x_3 — объем произведенной продукции, млн руб.:

№ п.п.	x_1	x_2	x_3
1	130	90	165
2	80	75	90
3	140	80	100
4	70	76	80
5	60	66	110
Среднее	96	77,4	109
Среднеквадратическое отклонение	36,5	8,7	33,2

Решение.

Нормируем исходные данные вычитанием среднего и делением на среднеквадратическое отклонение. Получим нормированные переменные:

№ п.п.	z_1	z_2	z_3
1	0,93	1,45	1,69
2	-0,44	-0,28	-0,57
3	1,21	0,30	-0,27
4	-0,71	-0,16	-0,87
5	-0,99	-1,31	0,03

Классификацию проведем при помощи иерархического агломеративного метода. Для построения матрицы расстояний воспользуемся евклидовым расстоянием. Например, расстояние между первым и вторым объектами:

$$d_{12}=\left[\left(0{,}93-(-0{,}44)\right)^2+\left(1{,}45-(-0{,}28)\right)^2+\left(1{,}69-(0{,}57)\right)^2\right]^{1/2}=3{,}15.$$

Первоначальная матрица расстояний D_0 характеризует расстояния между отдельными объектами, каждый из которых на первом шаге является отдельным кластером:

$$D_0=\begin{matrix}S_1\\S_2\\S_3\\S_4\\S_5\end{matrix}\begin{matrix}\begin{matrix}S_1 & S_2 & S_3 & S_4 & S_5\end{matrix}\\\begin{bmatrix}0 & 3{,}15 & 2{,}29 & 3{,}44 & 3{,}75\\ & 0 & 1{,}77 & 0{,}42 & 1{,}32\\ & & 0 & 2{,}06 & 2{,}74\\ & & & 0 & 1{,}49\\ & & & & 0\end{bmatrix}\end{matrix}.$$

Наиболее близкими являются объекты 2 и 4, расстояние между которыми 0,42. Объединим их в один кластер S_4.

Воспользуемся принципом «дальнего соседа» для вычисления расстояния между объектами. Тогда, например, между первым объектом и S_4: $d_{S_1S_4}=\max\{d_{12},d_{14}\}=\max\{3{,}15;3{,}44\}=3{,}44$ и т.д.

Получим матрицу расстояний D_1: $D_1=\begin{matrix}S_1\\S_3\\S_4\{2,4\}\\S_5\end{matrix}\begin{matrix}\begin{matrix}S_1 & S_3 & S_4\{2,4\} & S_5\end{matrix}\\\begin{bmatrix}0 & 2{,}29 & 3{,}44 & 3{,}75\\ & 0 & 2{,}06 & 2{,}74\\ & & 0 & 1{,}49\\ & & & 0\end{bmatrix}\end{matrix}$.

В матрице D_1 опять находим самые близкие кластеры. По скольку d_{45}=1,49, то объединяем кластеры S_4 и S_5, получим новый кластер S_4, содержащий объекты 2,4 и 5. Пересчитываем

расстояния $d_{S_1S_4}=\max\{d_{1IV},d_{15}\}=\max\{3{,}44;3{,}75\}=3{,}75$, $d_{S_3S_4}=\max\{d_{3IV},d_{35}\}=\max\{2{,}06;2{,}74\}=2{,}74$ и получаем новую матрицу D_2: $D_2=\begin{bmatrix}0 & 2{,}29 & 3{,}75\\ & 0 & 2{,}74\\ & & 0\end{bmatrix}$.

Объединяем кластеры S_1 и S_3, получим новый кластер S_1, содержащий объекты 1 и 3. Имеем два кластера $S_1\{1,3\}$ и $S_4\{2,4,5\}$: $d_{S_1S_4}=\max\{d_{1IV},d_{3IV}\}=\max\{3{,}75;2{,}74\}=3{,}75$ и $D_3=\begin{bmatrix}0 & 3{,}75\\ & 0\end{bmatrix}$.

На последнем шаге объединяем кластеры S_1 и S_4. Представим результаты классификации в виде дендрограммы на рис. 5.5.

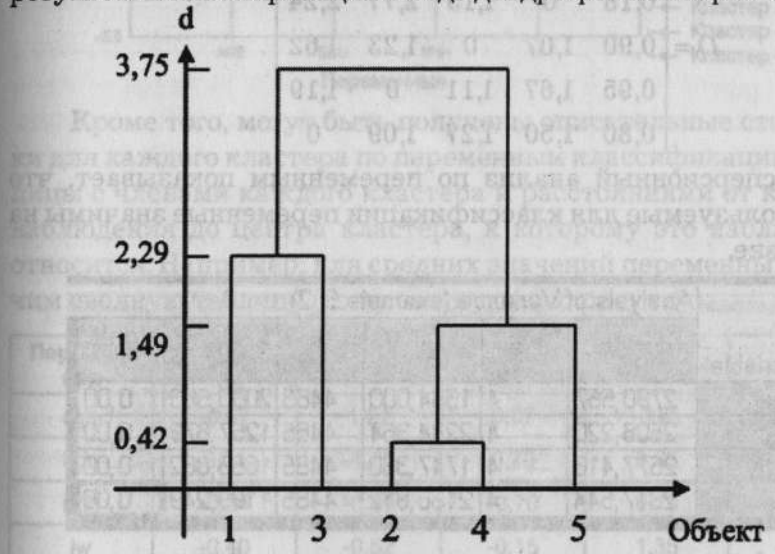


Рис. 5.5. Дендрограмма кластеризации пяти объектов

Пример 5.2. Исходные данные из массива msm.sta. Файл содержит подвыборку из 4490 наблюдений. Описание переменных: *lw* — логарифм заработной платы, *edu* — число лет образования, *nhh* — число членов домохозяйства, *age* — возраст главы домохозяйства. Необходимо классифицировать наблюдения на 5 групп.

Решение.

Воспользуемся пакетом прикладных программ Statistica и методом k-средних. Нормируем данные. В качестве начального расстояния выберем опцию: Sort distances and take observations at constant intervals, которая означает сортировку расстояний между объектами и выбор в качестве начальных центров кластеров объектов с постоянными расстояниями. После семи итераций получим следующие результаты.

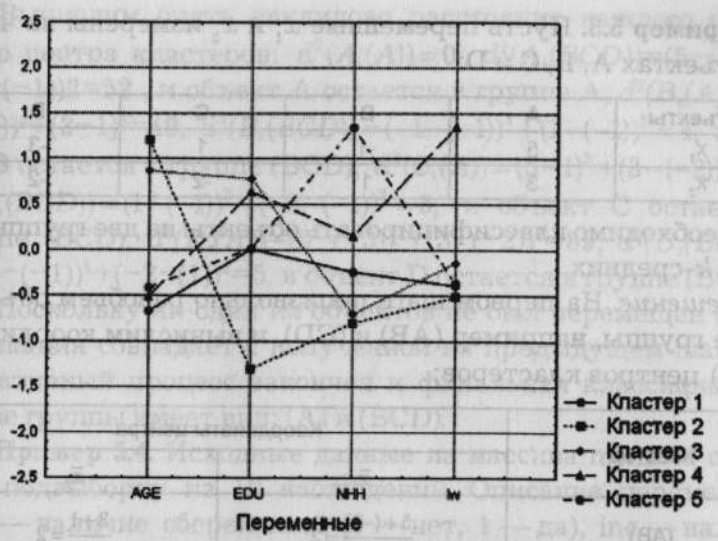
Матрица евклидовых расстояний (под диагональю) и квадратов расстояний (над диагональю) между кластерами:

$$D = \begin{array}{c} \begin{matrix} S_1 & S_2 & S_3 & S_4 & S_5 \end{matrix} \\ \begin{bmatrix} 0 & 0{,}39 & 0{,}81 & 0{,}91 & 0{,}64 \\ 0{,}18 & 0 & 1{,}15 & 2{,}77 & 2{,}24 \\ 0{,}90 & 1{,}07 & 0 & 1{,}23 & 1{,}62 \\ 0{,}95 & 1{,}67 & 1{,}11 & 0 & 1{,}19 \\ 0{,}80 & 1{,}50 & 1{,}27 & 1{,}09 & 0 \end{bmatrix} \end{array}.$$

Дисперсионный анализ по переменным показывает, что все используемые для классификации переменные значимы на 1% уровне.

	Analysis of Variance (example 5_2)					
Variable	Between SS	df	Within SS	df	F	signif. p
age	2790,557	4	1564,000	4485	2000,583	0,00
edu	2506,220	4	2234,354	4485	1257,679	0,00
nhh	2577,416	4	1747,360	4485	1653,882	0,00
lw	2317,544	4	2166,812	4485	1199,249	0,00

Информативен график средних значений переменных по кластерам:



Кроме того, могут быть получены описательные статистики для каждого кластера по переменным классификации и таблицы с членами каждого кластера и расстояниями от каждого наблюдения до центра кластера, к которому это наблюдение относится. Например, для средних значений переменных получим сводную таблицу (соответствующую рисунку):

Переменная	Кластер 1	Кластер 2	Кластер 3	Кластер 4	Кластер 5
age	-0,67	1,21	0,87	-0,51	-0,42
edu	0,01	-1,30	0,78	0,63	0,02
nhh	-0,24	-0,80	-0,70	0,15	1,34
lw	-0,40	-0,52	-0,15	1,35	-0,37

Интерпретация полученных результатов основывается на информации о членах кластеров и средних значениях переменных. Например, для кластера 2 получаем, что он содержит индивидов старших (пенсионных) возрастов с низким уровнем заработной платы, а кластер 4 — семейных индивидов молодых возрастов, имеющих высшее образование и высокую заработную плату.

Пример 5.3. Пусть переменные x_1 и x_2 измерены на четырех объектах A, B, C и D:

Объекты	A	B	C	D
x_1	5	-1	1	-3
x_2	3	1	-2	-2

Необходимо классифицировать объекты на две группы методом k-средних.

Решение. На первом шаге произвольно разобьем объекты на две группы, например (AB) и (CD), и вычислим координаты $(\bar{x}_1, \bar{x}_2)$ центров кластеров:

Кластер	Координаты центра	
	$\bar{x}_1$	$\bar{x}_2$
(AB)	$\frac{5+(-1)}{2}=2$	$\frac{3+1}{2}=2$
(CD)	$\frac{1+(-3)}{2}=-1$	$\frac{-2+(-2)}{2}=-2$

На втором шаге вычислим евклидово расстояние каждого объекта до центра кластеров и перераспределим каждый объект в ближайшую группу. Вычислим квадраты расстояний:

$d^2(A,(AB))=(5-2)^2+(3-2)^2=10$,

$d^2(A,(CD))=(5+1)^2+(3+2)^2=61$.

Таким образом, объект A остается в кластере (AB), поскольку расстояние от него до центра кластера наименьшее.

Далее $d^2(B,(AB))=(-1-2)^2+(1-2)^2=10$,

$d^2(B,(CD))=(-1+1)^2+(1+2)^2=9$.

Объект B ближе к кластеру (CD) и перемещается в этот кластер. Для групп (A) и (BCD) пересчитаем координаты центров:

Кластер	Координаты центра	
	$\bar{x}_1$	$\bar{x}_2$
(A)	5	3
(BCD)	$\frac{-1+1+(-3)}{3}=-1$	$\frac{1-2+(-2)}{3}=-1$

Вычислим опять евклидово расстояние каждого объекта до центра кластеров: $d^2(A,(A))=0$, $d^2(A,(BCD))=(5-(-1))^2+$ $+(3-(-1))2=52$, и объект A остается в группе A, $d^2(B,(A))=(5-$ $-(-1))^2+(3-1)^2=40$, $d^2(B,(BCD))=(-1-(-1))^2+(1-(-1))^2=4$, и объект B остается в группе (BCD), $d^2(C,(A))=(5-1)^2+(3-(-2))^2=41$, $d^2(C,(BCD))=(1-(-1))^2+(-2-(-1))^2=5$, и объект C остается в группе (BCD), $d^2(D,(A))=(5-(-3))^2+(3-(-2))^2=89$, $d^2(D,(BCD))=$ $=(-3-(-1))^2+(-2-(1))^2=5$ и объект D остается в группе (BCD).

Поскольку ни один из объектов не был перемещен (классификация совпадает с полученной на предыдущем шаге), то итеративный процесс закончен и финальная классификация на две группы имеет вид: (A) и (BCD).

Пример 5.4. Исходные данные из массива msm.sta содержат подвыборку из 10 наблюдений. Описание переменных: save — наличие сбережений (0 — нет, 1 — да), inc — наличие доходов (0 — нет, 1 — да), prop — имущественная обеспеченность (0 — нет, 1 — да), house — наличие жилья (0 — нет, 1 — да), catt —наличие скота (0 — нет, 1 — да), land — наличие земли в собственности (0 — нет, 1 — да). Необходимо классифицировать наблюдения на 3 группы.

№ п.п.	save	inc	prop	house	catt	land
1	0	1	1	0	0	0
2	1	1	1	1	1	1
3	0	1	1	0	0	0
4	0	1	0	0	0	0
5	1	1	1	1	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	1	1	0	0	0
9	0	0	0	1	0	0
10	0	1	1	0	0	1

Решение. Мы имеем случай, когда все переменные измерены в номинальной шкале. Чтобы провести кластерный анализ,

необходимо получить сначала матрицу сходства объектов, а затем перейти к матрице расстояний. Воспользуемся в качестве меры сходства коэффициентом ассоциативности Жаккара.

Коэффициент ассоциативности вычисляется по таблице сопряженности 2×2, в которой 1 указывает на наличие переменной и 0 на ее отсутствие

	1	0
1	a	b
0	c	d

Коэффициент Жаккара определяется как $J = a/(a+b+c)$ и изменяется от 0 до 1. Коэффициент Жаккара принимает в расчет лишь те признаки, которые характерны хотя бы для одного из объектов.

Например, для 1-го и 2-го объектов:

	1	0
1	2	0
0	4	0

и $J_{12}=2/(2+0+4)=0{,}33$, аналогично, например, для 2-го и 5-го объектов:

	1	0
1	4	2
0	0	0

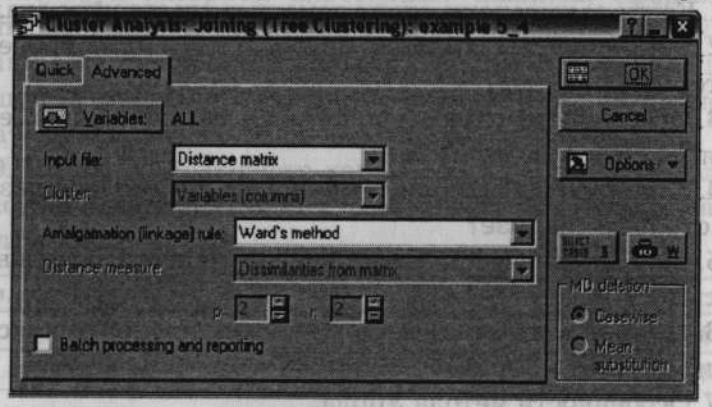
и $J_{25}=4/(4+2+0)=0{,}67$ и т.п., получим в результате для всех объектов матрицу сходства в виде:

	1	2	3	4	5	6	7	8	9	10
1	0	0,33	1	0,5	0,5	0	0	1	0	0,67
2	0,33	0	0,33	0,17	0,667	0	0	0,33	0,17	0,5
3	1	0,33	0	0,5	0,5	0	0	1	0	0,67
4	0,5	0,17	0,5	0	0,25	0	0	0,5	0	0,33
5	0,5	0,67	0,5	0,25	0	0	0	0,5	0,25	0,4
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	1	0,33	1	0,5	0,5	0	0	0	0	0,67
9	0	0,17	0	0	0,25	0	0	0	0	0
10	0,67	0,5	0,67	0,33	0,4	0	0	0,67	0	0

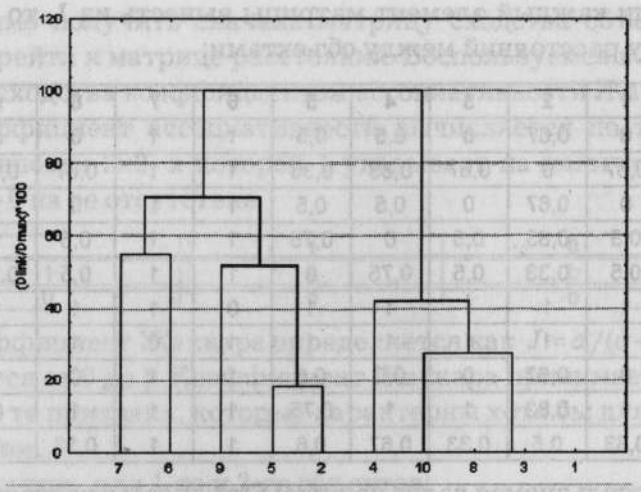
Если каждый элемент матрицы вычесть из 1, то получим матрицу расстояний между объектами:

	1	2	3	4	5	6	7	8	9	10
1	0	0,67	0	0,5	0,5	1	1	0	1	0,33
2	0,67	0	0,67	0,83	0,33	1	1	0,67	0,83	0,5
3	0	0,67	0	0,5	0,5	1	1	0	1	0,33
4	0,5	0,83	0,5	0	0,75	1	1	0,5	1	0,67
5	0,5	0,33	0,5	0,75	0	1	1	0,5	0,75	0,6
6	1	1	1	1	1	0	1	1	1	1
7	1	1	1	1	1	1	0	1	1	1
8	0	0,67	0	0,5	0,5	1	1	0	1	0,33
9	1	0,83	1	1	0,75	1	1	1	0	1
10	0,33	0,5	0,33	0,67	0,6	1	1	0,33	1	0

Воспользуемся далее пакетом Statistica и в качестве исходной информации для кластерного анализа вычисленной матрицей расстояний. Классификацию проведем методом Уорда.



Получим дендрограмму, представленную на рисунке. Видно, что объекты 1, 3 и 8 не различимы. Поскольку по условию задачи k=3, то получаем разбиение: (1, 3, 4, 8, 10), (6, 7) и (2, 5, 9). При этом замечаем, что по исходным данным объекты 6 и 7 также одинаковы, а 2 и 5 почти схожи.



Вопросы и задачи

1. Какие задачи решаются с помощью кластерного анализа?

2. Какие меры сходства используются при проведении кластерного анализа?

3. Особенности параметрической классификации без обучения.

4. Какие меры расстояний между объектами используются в кластерном анализе?

5. Как оценивается качество полученного разбиения на классы?

6. Принцип «работы» иерархических процедур классификации.

7. Особенности метода Уорда.

8. Алгоритм метода k-средних.

9. По 15 фирмам имеются следующие данные

№ п.п.	Фондовооруженность труда, млн руб./чел.	Фондоотдача основных фондов, руб./руб.	Удельный вес рабочих в составе персонала
1	4,82	1,67	0,46
2	3,85	1,78	0,72

№ п.п.	Фондовооруженность труда, млн руб./чел.	Фондоотдача основных фондов, руб./руб.	Удельный вес рабочих в составе персонала
3	4,75	1,23	0,67
4	5,35	1,32	0,71
5	8,7	0,75	0,66
6	7,3	1,15	0,72
7	6,4	1,26	0,7
8	5,9	1,43	0,75
9	6	1,28	0,63
10	8,95	0,95	0,76
11	7,41	1,18	0,69
12	4,71	1,9	0,71
13	5,03	1,81	0,72
14	6,94	1,29	0,73
15	7,95	0,98	0,7

Используя алгоритм кластерного анализа, сформируйте из первых десяти наблюдений две обучающие выборки. На основании полученных выборок проведите классификацию 5 оставшихся фирм. Дайте экономическую интерпретацию результатов дискриминантного анализа.

10. Шесть домохозяйств характеризуются показателями: x_1 — потребление фруктов (кг/месяц) и x_2 — потребление молока (л/месяц):

№ п.п.	1	2	3	4	5	6
x_1	21,4	16,5	9,7	18,2	6,6	8,0
x_2	8,1	4,2	5,5	9,4	7,5	5,7

Требуется: с помощью иерархического агломеративного алгоритма провести классификацию этих предприятий и построить дендрограмму:

1) при использовании обычной евклидовой метрики — методом: а) ближайшего соседа, б) дальнего соседа, в) центра тяжести, г) средней связи;

2) при использовании взвешенной евклидовой метрики (с весами 0,2 и 0,8) методом ближайшего соседа.

Задание к лабораторному практикуму
Время выполнения — 4 часа

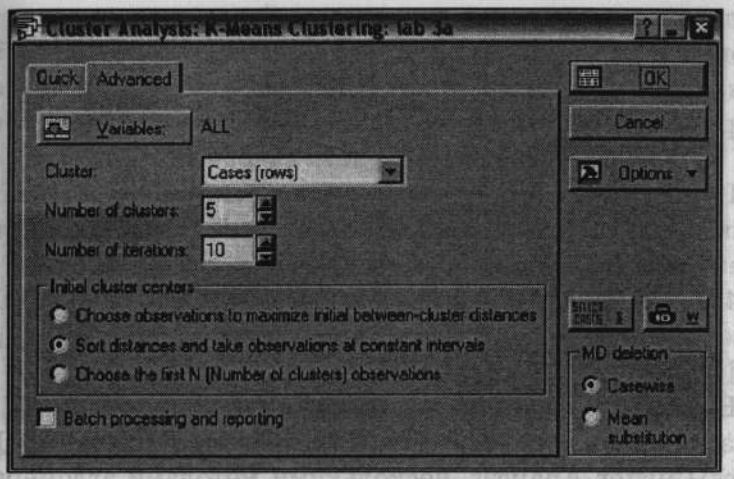
1. Запустите приложение *Statistica*. Откройте файл *Lab3.sta*, воспользовавшись меню File\Open. Файл содержит подвыборку из массива msm.sta по индивидам. Описание переменных: *lw* — логарифм заработной платы, *edu* — число лет образования, *nhh* — число членов домохозяйства, *dd* — доля доходов главы домохозяйства в семейном бюджете, *age* — возраст, *pm* — процент заработков, который дает основная работа. Необходимо классифицировать наблюдения.

2. Выполните расчет описательных статистик по переменным выборки. Сделайте выводы. Почему нельзя использовать данные в натуральном виде? В файле *Lab3a.sta* содержатся стандартизованные переменные.

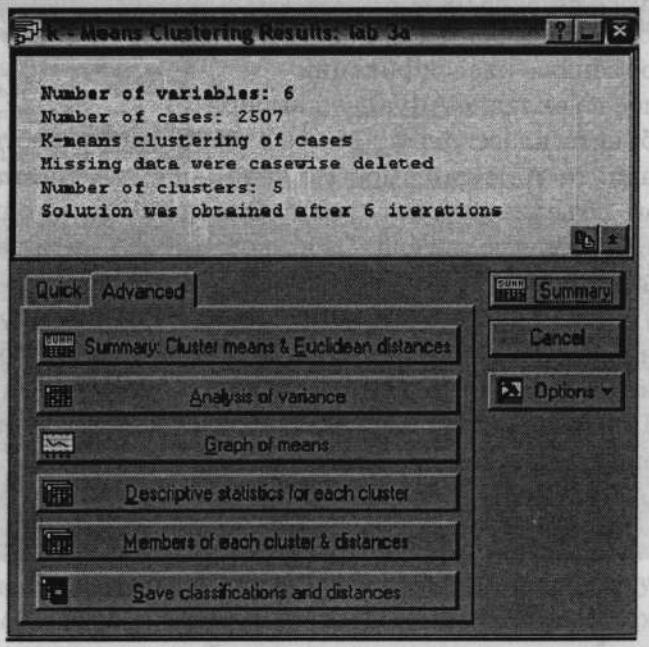
3. Выполните кластерный анализ имеющихся индивидов, воспользовавшись меню Statistics\Multivariate Exploratory Techniques\Cluster Analysis и методом автоматической классификации *k*-средних, указав в появившемся окне K-means clustering и нажав OK. В окне кластерного анализа методом *k*-средних необходимо указать переменные классификации (укажите все имеющиеся переменные). Далее во вкладке Advanced выберите объекты классификации — в меню Cluster укажите Cases; задайте число кластеров — например, три; обратите внимание на возможность выбора начальных центров кластеров (Initial cluster centers).

После запуска вычислительной процедуры появится окно результатов, в верхней части которого содержится общая информация о классификации и количество итераций, после которых найдено решение. Выберите вкладку Advanced. Опция Cluster Means&Euclidean Distances позволяет получить таблицы с средними по переменным для каждого кластера и расстояниями между кластерами.

Опция Analysis of variance позволяет выполнить дисперсионный анализ для проверки значимости для классификации каждой из использованных переменных.



Кнопка Graph of means позволяет просмотреть средние значения для каждого кластера на графике.



Кнопка Descriptive statistics for each cluster позволяет получить описательные статистики (математическое ожидание,

стандартное отклонение и дисперсию) для каждого кластера. Наконец опция Members of each cluster&distances дает объекты (наблюдения) каждого класса и расстояние от объектов до центра кластера, которому принадлежит этот объект. Кнопка Save позволяет сохранить результаты классификации.

Попробуйте изменить количество кластеров и состав переменных, по которым строится классификация. Как это влияет на результаты разбиения?

4. Организуйте случайную подвыборку наблюдений Data\ Subset и далее, выбирая все переменные, укажите в Simple Random Sampling 2% percent of cases.

Выполните кластерный анализ имеющихся данных, воспользовавшись меню Statistics\Multivariate Exploratory Techniques\Cluster Analysis, несколькими методами агломеративной классификации, указав в появившемся окне Joining (tree clustering) и нажав OK.

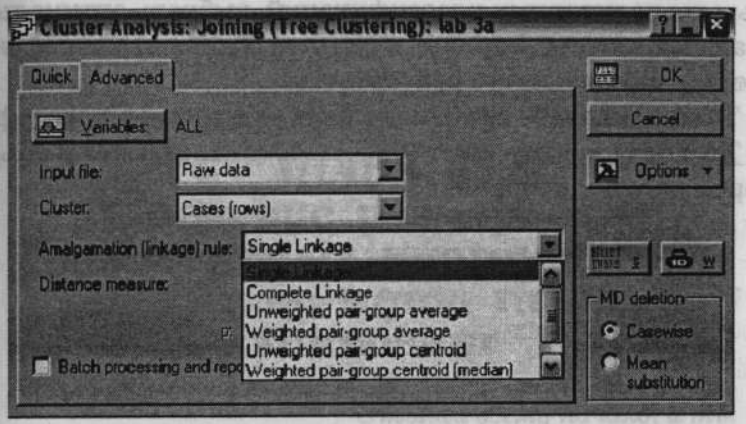
В появившемся окне кластерного анализа необходимо указать переменные классификации.

Далее во вкладке Advanced выберите:

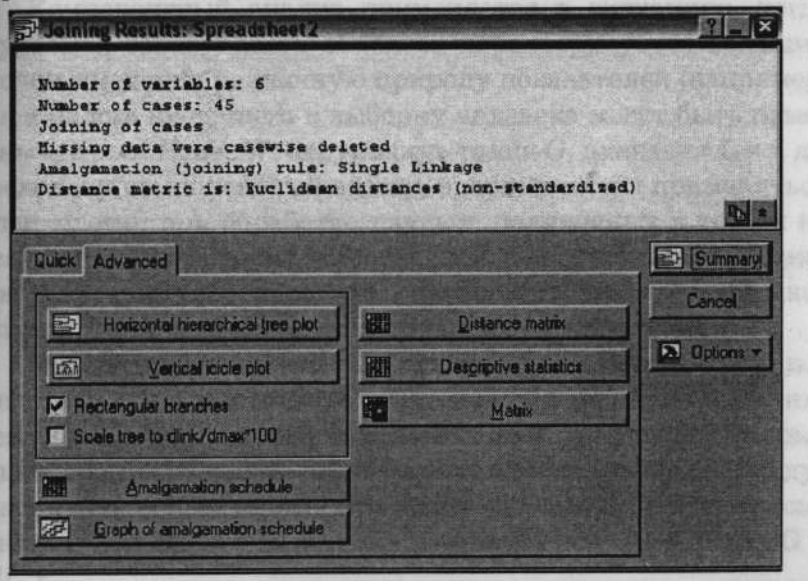
– объекты классификации — в меню Cluster укажите Cases;

– задайте Amalgamation (linkage) rule, т.е. метод иерархического объединения кластеров: Single linkage — одиночной связи (ближайшего соседа), Complete linkage — полной связи (дальнего соседа), Unweighted pair-group average — невзвешенный метод средней связи, Weighted pair-group average — взвешенный метод средней связи, Unweighted pair-group centroid — невзвешенный центроидный метод, Weighted pair-group centroid — взвешенный центроидный метод (медианной связи), Ward's method — метод Уорда;

– меру расстояния между объектами Distance measure.



После запуска вычислительной процедуры появится окно результатов, в верхней части которого содержится общая информация о классификации.



Нажав на кнопку Vertical hierarchical tree plot, получим иерархическое дерево результатов агломеративной классификации.

Рекомендуется выбирая различные правила объединения кластеров и меры расстояний между объектами, сравнить ре-

зультаты полученных классификаций, выбрать оптимальный, по вашему мнению, вариант и сделать содержательные экономические выводы по его результатам.

Также в пакете Statistica реализован метод классификации Two-way joining, в котором классифицируются и объекты и переменные одновременно.

ГЛАВА 6
Снижение размерности исследуемых многомерных признаков: метод главных компонент

Отыщи всему начало, и ты многое поймешь.

К. Прутков. Мысли и афоризмы

Компонентный анализ применяется в ситуациях, когда изучаемая совокупность объектов характеризуется большим числом имеющих различную природу показателей (например, для каждого входящего в выборку человека могут быть измерены его рост H, вес W, окружность талии G, длина ног L и т. д.). Поскольку компонентный анализ впервые начал применяться в психологии при обработке данных, полученных в тестах по выявлению умственных способностей, различные характеристики, относящиеся к исходным данным, в литературе часто называют «тестовыми переменными», или параметрами.

Метод предназначен для структуризации данных посредством сведения множества тестовых переменных к меньшему числу переменных (компонент или факторов), которые объясняли бы бульшую часть вариации в значениях исследуемых данных. Компонента представляет собой линейную комбинацию исходных переменных: компонента $= aH + bW + cG + dL + \ldots$.

Требуется определить числовые значения коэффициентов a, b, c, d при условии, что выполняется требование о равенстве единице дисперсии каждой из компонент. Поскольку априорная информация о значениях компонент отсутствует, то с по-

мощью специальной аналитической техники в рамках компонентного анализа производятся преобразования, приводящие к тому, что p переменных выражают через главные компоненты, а затем определяют их величину и содержательное значение. Важным свойством компонент является то, что каждая из них по порядку учитывает максимум суммарной дисперсии параметров. То есть первая главная компонента есть линейная комбинация исходных параметров, учитывающая максимум их суммарной дисперсии, вторая главная компонента не коррелирует с первой и учитывает максимум оставшейся дисперсии и т.д. до тех пор, пока вся дисперсия не будет учтена. Сумма дисперсий всех главных компонент равна сумме дисперсий всех исходных параметров.

Основными типами решаемых методом главных компонент задач являются:

- отыскание скрытых, но объективно существующих закономерностей, определяемых воздействием внутренних и внешних причин,
- описание изучаемого процесса числом главных компонент, значительно меньшим, чем число первоначально взятых признаков,
- выявление и изучение стохастической связи признаков с главными компонентами,
- возможность использования полученных результатов для прогнозирования процесса на основе построения регрессии.

Метод главных компонент по своей сути позволяет для пространства, образуемого p признаками, перейти к меньшему числу k признаков — главных компонент, причем $k<p$. Различия между объектами зависят от доли вариации, связанной (объясняемой) с данной главной компонентой. Предполагается, что каждому признаку свойственна факторная структура, связи между признаками и факторами (главными компонентами) линейны, для данного признака эффект воздействия факторов суммируется.

Переход к меньшему числу k признаков (главных компонент) позволяет уменьшить объем информации и содержательно проинтерпретировать новые признаки (главные компоненты).

Пусть имеется p случайных величин $x_1, x_2, ..., x_p$, обозначим Σ ковариационную матрицу этих случайных величин

$$\Sigma = \begin{bmatrix} V(x_1) & \operatorname{cov}(x_1, x_2) & \dots & \operatorname{cov}(x_1, x_p) \\ \operatorname{cov}(x_2, x_1) & V(x_2) & & \vdots \\ \operatorname{cov}(x_p, x_1) & \dots & \dots & V(x_p) \end{bmatrix}$$ и обозначим ρ корреляционную матрицу переменных $x_1, x_2, ..., x_p$.

Заметим, что в общем случае переменные $x_1, x_2, ..., x_p$ могут не иметь нормального распределения.

Обозначим вектор переменных $\mathbf{X}^T=(x_1, x_2, ..., x_p)$. Запишем линейные комбинации

$$\begin{aligned} Y_1 &= a_1^T\mathbf{X} = a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p; \\ Y_2 &= a_2^T\mathbf{X} = a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p; \\ &\dots \\ Y_p &= a_p^T\mathbf{X} = a_{p1}x_1 + a_{p2}x_2 + \dots + a_{pp}x_p. \end{aligned} \qquad (6.1)$$

В (6.1) дисперсия: $V(Y_i) = a_i^T \Sigma a_i, i = 1, ..., p$ (6.2)

и ковариация $\operatorname{cov}(Y_i, Y_j) = a_i^T \Sigma a_j, i = 1, ..., p$.

Если переменные $Y_1, ... Y_p$ имеют наибольшую дисперсию (6.2) и попарно независимы, то эти переменные $Y_1, ... Y_p$ соответствуют *главным компонентам*.

Первая *главная компонента* является линейной комбинацией с максимальной дисперсией, т.е. числа $a_{11}, ..., a_{1p}$ в линейной комбинации $a_1^T\mathbf{X}$ выбираются из условия максимизации дисперсии $V(a_1^T\mathbf{X})$ при выполнении ограничения $a_1 a_1 = 1$.

Вторая *главная компонента* есть линейная комбинация $a_2^T\mathbf{X}$ при условии максимизации дисперсии $V(a_2^T\mathbf{X})$, с учетом ограничений $a_2^T a_2 = 1$ и $\operatorname{cov}(a_1^T\mathbf{X}, a_2^T\mathbf{X}) = 0$ и т.п.

i-я главная компонента есть линейная комбинация $a_i^T\mathbf{X}$ при условии максимизации дисперсии $V(a_i^T\mathbf{X})$ и выполнения ограничений $a_i^T a_i = 1$ и $\operatorname{cov}(a_i^T\mathbf{X}, a_j^T\mathbf{X}) = 0, i \neq j, j < i, j = 1, ..., p$.

Теорема. Пусть Σ — ковариационная матрица, соответствующая случайному вектору $\mathbf{X}^T = (x_1, ..., x_p)$, и пусть Σ имеет собственные значения $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_p \geq 0$ и соответствующие им собственные векторы $e_1, e_2, ..., e_p$, тогда i-я главная компонента определяется как

$$Y_i = e_i^T \mathbf{X} = e_{i1}x_1 + e_{i2}x_2 + ... + e_{ip}x_p. \qquad (6.3)$$

В этом случае

$$V(Y_i) = e_i^T \Sigma e_i = \lambda_i, i = 1, ..., p, \quad \text{cov}(Y_i, Y_j) = e_i^T \Sigma e_j = 0, \quad i \neq j.$$

Если некоторые λ_i одинаковы, то выбор соответствующих e_i и Y_i не является единственным.

Согласно теореме главные компоненты являются некоррелированными и имеют дисперсии, равные собственным числам ковариационной матрицы Σ.

Теорема. Пусть $\mathbf{X}^T = (x_1, ..., x_p)$ имеет ковариационную матрицу Σ с собственными значениями $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_p \geq 0$ и соответствующими им собственными векторами $e_1, e_2, ..., e_p$ и пусть определены главные компоненты: $Y_1 = e_1^T \mathbf{X}, Y_2 = e_2^T \mathbf{X}, ..., Y_p = e_p^T \mathbf{X}$.

Тогда $\sigma_{11}^2 + \sigma_{22}^2 + ... + \sigma_{pp}^2 = \sum_{i=1}^{p} V(\mathbf{X}_i) = \lambda_1 + \lambda_2 + ... + \lambda_p = \sum_{i=1}^{p} V(Y_i)$,

где σ_{ii}^2 — элементы главной диагонали (дисперсии) ковариационной матрицы Σ.

Таким образом, теорема утверждает, что суммарная величина дисперсии генеральной совокупности имеющихся наблюдений $x_1, x_2, ..., x_p$ при переходе к новым переменным (главным компонентам) не изменяется.

Можно найти долю дисперсии, объясненную i-й главной компонентой в общей сумме дисперсии: $\omega_i = \dfrac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$.

Если большая часть общей дисперсии (например, 70–90%) объясняется одной, двумя или тремя главными компонентами, то эти компоненты могут быть использованы вместо исходных p переменных $x_1, x_2, ..., x_p$ без существенной потери информации.

Таким образом суть метода главных компонент заключается в том, чтобы от исходных p случайных величин перейти к меньшему числу переменных, объясняющих большую часть дисперсии (80-90% или хотя бы 70%) исходных переменных.

Теорема. Если $Y_1 = e_1^T \mathbf{X}, Y_2 = e_2^T \mathbf{X}, ..., Y_p = e_p^T \mathbf{X}$ являются главными компонентами, полученными на основе ковариационной матрицы Σ, то коэффициент корреляции между Y_i и x_k равен

$$r_{Y_i X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, i, k = 1, 2, ..., p. \qquad (6.4)$$

где λ_i — собственные числа, e_i — собственные векторы матрицы Σ.

Таким образом (6.4) показывает, что величина e_{ik} измеряет важность k-й переменной в i-й главной компоненте.

Главные компоненты могут быть получены для многомерных нормально распределенных переменных. Предположим, что наблюдения $X_1, ..., X_n$ имеют нормальное распределение с параметрами $\boldsymbol{\mu}$ и Σ: $N_p(\boldsymbol{\mu}, \Sigma)$. Тогда могут быть вычислены главные компоненты $Y_1 = e_1^T \mathbf{X}, Y_2 = e_2^T \mathbf{X}, ..., Y_p = e_p^T \mathbf{X}$ и оказывается справедливым равенство $(X - \boldsymbol{\mu})\Sigma^{-1}(X - \boldsymbol{\mu}) = c^2$,

где $c^2 = \mathbf{X}^T \Sigma^{-1} \mathbf{X} = \dfrac{1}{\lambda_1} Y_1^2 + ... + \dfrac{1}{\lambda_p} Y_p^2$.

Геометрически переход к главным компонентам показан на рис. 6.1. Таким образом, в случае нормальности переменных переход к главным компонентам означает поворот системы координат на некоторый угол θ при неизменной плотности распределения случайных величин внутри некоторого эллипса.

Главные компоненты могут быть получены также для стандартизованных переменных:

$$\mathbf{Z} = \left(\mathbf{V}^{1/2}\right)^{-1}(X - \boldsymbol{\mu}) \Leftrightarrow Z_1 = \frac{x_1 - \mu_1}{\sqrt{\sigma_{11}^2}}, Z_2 = \frac{x_2 - \mu_2}{\sqrt{\sigma_{22}^2}}, ..., Z_p = \frac{x_p - \mu_p}{\sqrt{\sigma_{pp}^2}},$$

где μ_i — математическое ожидание и σ_{ii}^2 — дисперсия, вычисленные по выборке для соответствующих переменных, $\mathbf{V}$ — матрица стандартных отклонений.

Ковариационная матрица $\mathrm{cov}(Z)=(V^{1/2})^{-1}\Sigma(V^{1/2})^{-1}=\boldsymbol{\rho}$, где $\boldsymbol{\rho}$ — корреляционная матрица переменных $\mathbf{X}$. Тогда главные компоненты: $Y_i=\mathbf{e}_i^T Z=\mathbf{e}_i^T(V^{1/2})^{-1}(X-\boldsymbol{\mu})$, $i=1,\ldots,p$ и, кроме того, $\sum_{i=1}^{p}V(Y_i)=\sum_{i=1}^{p}V(Z_i)=\boldsymbol{\rho}$ и $\rho_{Y_iZ_k}=e_{ik}\sqrt{\lambda_i}$, $i,k=1,2,\ldots,p$. Здесь e_l — собственные векторы матрицы $\boldsymbol{\rho}$ и λ_i — собственные числа матрицы $\boldsymbol{\rho}$.

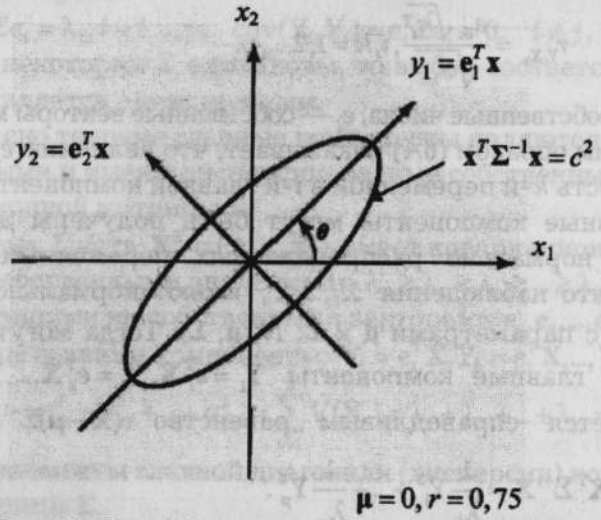


Рис. 6.1. *Переход к главным компонентам в случае нормального распределения переменных*

Алгебраические преобразования исходных данных X позволяют выделить главные компоненты Y и установить их пространственное расположение. Задача интерпретации главных компонент и определения для них названий решается затем субъективно на основе весовых коэффициентов a_{ij} из (6.1). Для каждой главной компоненты Y множество значений a_{ij} условно разбивается на четыре подмножества с нечеткими границами [9]:

– подмножество незначимых весовых коэффициентов,

– подмножество значимых весовых коэффициентов (на практике считается, что коэффициент значим, если его значение превышает 0,6 по модулю),

– подмножество значимых весовых коэффициентов, не участвующих в формировании названия главной компоненты,

– подмножество значимых весовых коэффициентов, участвующих в формировании названия главной компоненты.

Подтверждение значимости признаков x, участвующих в формировании названия компоненты, можно получить расчетным путем при определении коэффициента информативности: $K_u=\sum_{i=1}^{m}a_{ij}^2\Big/\sum_{i=1}^{p}a_{ij}^2$, где в числителе суммирование осуществляется по участвующим в интерпретации коэффициентам. Считается удовлетворительным, если K_u не ниже 0,75.

Пример 6.1. Деятельность предприятий характеризуется следующими показателями

№ п.п.	Трудоемкость единицы продукции, x_1	Удельный вес покупных изделий, x_2	Коэффициент сменности оборудования, x_3	Индекс снижения себестоимости продукции, q
1	0,51	0,20	1,47	21,9
2	0,36	0,64	1,27	48,4
3	0,23	0,42	1,51	173,5
4	0,26	0,27	1,46	74,1
5	0,27	0,37	1,27	68,6
6	0,29	0,38	1,43	60,8
7	0,01	0,35	1,50	355,6
8	0,02	0,42	1,35	264,8
9	0,18	0,32	1,41	526,6
10	0,25	0,33	1,47	118,6

Приняв за результативный признак q, построить уравнение регрессии на главные компоненты, наиболее тесно связанные с q. Дать экономическую интерпретацию результатов.

Решение.

Перейдем к стандартизованным переменным, вычитая среднее и деля на стандартное отклонение для каждой из x:

№ п.п.	Трудоемкость единицы продукции, z_1	Удельный вес покупных изделий, z_2	Коэффициент сменности оборудования, z_3
1	1,84	-1,46	0,63
2	0,83	2,32	-1,62

№ п.п.	Трудоемкость единицы продукции, z_1	Удельный вес покупных изделий, z_2	Коэффициент сменности оборудования, z_3
3	-0,05	0,43	1,08
4	0,15	-0,86	0,52
5	0,22	0,00	-1,62
6	0,35	0,09	0,18
7	-1,54	-0,17	0,97
8	-1,48	0,43	-0,72
9	-0,39	-0,43	-0,05
10	0,08	-0,34	0,63

Получим корреляционную матрицу:

$$\rho = \begin{bmatrix} 1 & -0,125 & -0,090 \\ -0,125 & 1 & -0,593 \\ -0,090 & -0,593 & 1 \end{bmatrix}.$$

Характеристическое уравнение для корреляционной матрицы получаем как $\det(p) = \begin{vmatrix} 1-\lambda & -0,125 & -0,090 \\ -0,125 & 1-\lambda & -0,593 \\ -0,090 & -0,593 & 1-\lambda \end{vmatrix} =$

$$= (1-\lambda)[(1-\lambda)^2 - 0,593^2] - (-0,125)[-0,125(1-\lambda) - (-0,593)(-0,090)] +$$
$$+(-0,090)[(-0,125)(-0,593) - (-0,090)(1-\lambda)] = -\lambda^3 + 3\lambda^2 - 2,625\lambda +$$
$$+0,611 = 0.$$

Решения кубического уравнения суть: $\lambda_1 = 1,594$, $\lambda_2 = 1,036$, $\lambda_3 = 0,370$.

Убеждаемся, что сумма $\sum_i \lambda_i = 3$. Также имеем, что первая главная компонента объясняет $1,594/3 \cdot 100\% = 53\%$ вариации, вторая 35%, третья — 12%. Первые две главные компоненты объясняют таким образом 88% всей дисперсии переменных.

Собственный вектор, соответствующий λ_1 и первой главной компоненте, находим из системы:

$$\begin{cases} (1-1,594)e_{11} - 0,125e_{12} - 0,09e_{13} = 0; \\ -0,125e_{11} + (1-1,594)e_{12} - 0,593e_{13} = 0; \\ -0,09e_{11} - 0,593e_{12} + (1-1,594)e_{13} = 0. \end{cases}$$

Решая, получаем $e_1 = (e_{11}\, e_{12}\, e_{13}) = (0,043\ -0,710\ 0,703)$. Аналогично $e_2 = (0,971\ -0,136\ -0,197)$ и $e_3 = (0,235\ 0,691\ 0,684)$. Непосредственно можно убедиться, что значения главных компонент нормированы — сумма квадратов значений каждой компоненты дает 1.

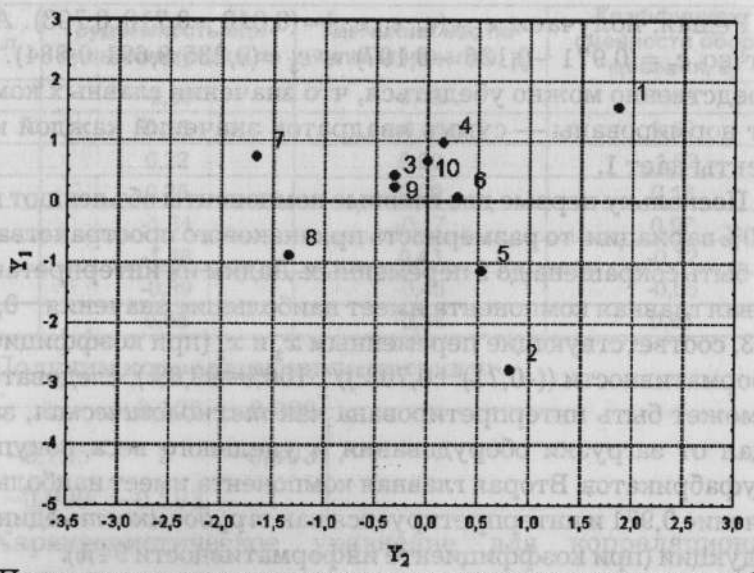
Поскольку первые две главные компоненты объясняют почти 90% вариации, то размерность признакового пространства может быть сокращена до 2 переменных. Дадим их интерпретацию. Первая главная компонента имеет наибольшие значения −0,71 и 0,703, соответствующие переменным x_2 и x_3 (при коэффициенте информативности $((-0,71)^2+0,703^2)/1 \cdot 100\% = 99,8\%$), следовательно, может быть интерпретирована как *технологическая*, зависящая от загрузки оборудования и удельного веса покупных полуфабрикатов. Вторая главная компонента имеет наибольшее значение 0,971 и интерпретируется как *трудоемкость* единицы продукции (при коэффициенте информативности 94%).

Значения главных компонент для каждого предприятия получаются по формуле (6.3) и приведены в таблице.

Например, значение Y_1 для первого наблюдения:
$0,043 \cdot 1,84 + (-0,71) \cdot (-1,46) + 0,703 \cdot 0,63 = 1,562$.

№ п.п.	Y_1	Y_2
1	1,562	1,864
2	-2,756	0,806
3	0,452	-0,324
4	0,982	0,160
5	-1,131	0,530
6	0,081	0,295
7	0,737	-1,667
8	-0,876	-1,350
9	0,257	-0,314
10	0,692	0,002

Распределение предприятий по значениям двух первых главных компонент показано на рисунке (построен по последней таблице). Ясно, что условно хорошие предприятия находятся в левой полуплоскости по значениям переменной Y_2.



Построим уравнение регрессии индекса снижения себестоимости продукции (предварительно стандартизовав его — $\tilde{q}$) на первые две главные компоненты: $\tilde{q} = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \varepsilon$. Получим $\tilde{q} = \underset{[0,45]}{0,096} Y_1 - \underset{[2,52]}{0,672} Y_2$, в квадратных скобках показаны значения t-критерия. При этом $R^2_{скорр} = 0,33$, $F(2,7) = 3,27$. Свободный член регрессии равен нулю. Первая главная компонента не значима, вторая — значима на 5% уровне, уравнение регрессии в целом значимо на 10% уровне. Можно упростить уравнение, удалив незначимую Y_1, при этом коэффициенты уравнения не изменятся (регрессоры не коррелированы): $\tilde{q} = -0,672 Y_2$.

Одна вторая главная компонента достаточно полно характеризует изменение индекса снижения себестоимости, причем наибольший вклад в формирование этой компоненты вносит переменная x_1. Перейдем к исходным (нестандартизованным) переменным в последнем уравнении:

$$\tilde{q} = -0,672 Y_2 = -0,672(0,971 z_1 - 0,136 z_2 - 0,197 z_3) \Rightarrow$$

$$\frac{q - 171,29}{163,76} = -\left(0,653 \frac{x_1 - 0,238}{0,148} - 0,091 \frac{x_2 - 0,370}{0,116} - 0,132 \frac{x_3 - 1,414}{0,089}\right)$$

и $q = -391,41 - 723,84 x_1 + 128,55 x_2 + 243,87 x_3$.

Пример 6.2.

Следующий простой пример не относится непосредственно к области экономики, но наглядность и ясность интерпретации полученного решения делают его полезным для понимания сути структуризации данных путем «сжатия» исходного признакового пространства*.

Имеются данные об измерениях шести характеристик физического развития большой группы новорожденных детей, принадлежащих некоторой выборке. Исходная информация для анализа сведена в таблицу коэффициентов парной корреляции всех исходных переменных.

Корреляция между шестью характеристиками физического развития группы детей

Показатель	H	L	A	C	G	W
Рост	1,0	0,7	0,9	0,3	0,2	0,8
Длина ног	0,7	1,0	0,8	0,4	0,3	0,6
Длина рук	0,9	0,8	1,0	0,4	0,5	0,7
Окружности груди	0,3	0,4	0,4	1,0	0,8	0,3
Окружность талии	0,2	0,3	0,5	0,8	1,0	0,4
Вес	0,8	0,6	0,7	0,3	0,4	1,0

Цель анализа состоит в проведении группировки или структуризации исходных переменных с помощью небольшого числа компонент.

Решение.

Поскольку число исходных переменных равно 6, то выделение компонент путем непосредственного решения системы уравнений подобно тому, как это представлено в примере 6.1, становится весьма громоздким и затруднительным для ручного счета. Для получения решения в подобном случае необходимы специальные вычислительные процедуры, реализованные в пакетах прикладных программ по статистике.

* *Эренберг А.* Анализ и интерпретация статистических данных. — М.: Финансы и статистика, 1981. — С.283–293.

Значения главных компонент

Переменные	Компоненты		
	1	2	3
Рост (H)	0,9	-0,4	0,1
Длина ног (L)	0,8	-0,2	-0,4
Длина рук (A)	0,9	-0,2	-0,1
Окружности груди (C)	0,6	0,7	-0,1
Окружность талии (G)	0,6	0,7	0,2
Вес (W)	0,8	-0,2	0,4
Сумма квадратов факторных нагрузок	3,6	1,3	0,4

Первая главная компонента объясняет 3,6/6 или 60%, общей дисперсии фактических данных. Вторая — примерно 22% и третья компонента — примерно 7%. Вообще говоря, при шести переменных может быть выделено шесть компонент. Поскольку первые компоненты оказались довольно существенными, роль трех оставшихся должна быть незначительной. Каждая из них в среднем сможет объяснить только около 4% общей дисперсии (тогда в сумме для всех шести переменных мы получим 100%).

Столь незначительными компонентами можно вполне пренебречь. Таким образом, в компонентном анализе число выделяемых компонент меньше количества переменных. На практике, когда число переменных достигает 20 и более, обычно выделяются 3, 4 или 5 компонент (они могут быть затем рассмотрены более детально), т.е. метод главных компонент приводит к существенному сокращению числа переменных.

Теперь, когда произведено выделение нескольких компонент с помощью метода главных компонент, следующий шаг анализа состоит в интерпретации полученных результатов. Главная компонента 1 объясняет 60% общей дисперсии фактических данных (80% дисперсии *H*, 64% дисперсии L и т. д.).

Интерпретация компоненты обычно проводится посредством сравнения между собой значений полученных факторных нагрузок, в результате чего становится ясным, какие из исходных переменных сильнее всего коррелируют с данной компонентой. В результате рассматриваемая компонента получает соответствующее название.

Так, все представленные в таблице коэффициенты корреляции с компонентой 1 имеют высокое значение и положительны по своей величине. Таким образом, эта компонента довольно сильно коррелирует со всеми шестью характеристиками физического развития: если значения этих шести характеристик у какого-либо человека велики, то значение первой компоненты будет большим. Следовательно, эту компоненту можно интерпретировать как «общий показатель» физического развития. Вторая компонента положительно коррелирует с окружностью груди и окружностью талии, но ее связь с ростом и другими характеристиками тела отрицательна. Таким образом, ее можно рассматривать в качестве показателя «формы» тела. Интерпретация третьей компоненты в этом смысле затруднительна.

Пример 6.3. Следующий пример основан на данных, характеризующих отраслевую структуру малых предприятий одного из регионов РФ за период с 1999 по 2001 гг.

Для выявления внутренней структуры данных осуществлен компонентный анализ.

Поскольку при проведении компонентного анализа желательно соблюдение единого масштаба данных, то признаки преобразованы в относительные величины: доля предприятий отрасли среди всех малых предприятий — первый признак, объем производства в расчете на одного работника отрасли — второй признак, среднее число работников на одно предприятие отрасли — третий признак, доля средней заработной платы по отрасли от средней заработной платы по всем предприятиям — четвертый признак.

В качестве наблюдений рассматривались отдельные отрасли. Поскольку интерес представляют и структурные изменения, происходившие в течение анализируемого периода, то анализ проведен отдельно по каждому году. По результатам расчетов в 1999 г.

получена факторная структура, включающая три компоненты, объясняющие 99,5% суммарной вариации признаков (см. табл.).

Значения главных компонент, 1999 год

№ компоненты	Доля предприятий отрасли	Объем производства в расчете на одного работника отрасли, тыс руб.	Среднее число работников на одном предприятии отрасли, чел.	Доля средней заработной платы по отрасли в процентах от средней по всем малым предприятиям	Процент объясняемой дисперсии
1	0,001	0,961	-0,227	0,971	48,08
2	0,984	-0,031	-0,196	0,046	25,78
3	-0,176	-0,204	0,954	-0,159	25,63

Первая компонента обусловлена дисперсией второго и четвертого признаков, её можно определить как относительную доходность функционирования малых предприятий в отраслях экономики. Структура нагрузки второй компоненты позволяет трактовать её как эффект масштаба отрасли: действительно, с одной стороны, налицо такая отрасль, как общественное питание и торговля, по которой доля малых предприятий составляет 34,4%, и с другой стороны, здравоохранение, физическая культура и социальное обеспечение, занимающие лишь 0,5 % от сегмента малых предприятий. Третья компонента характеризует отрасли по численности работников предприятий.

Факторная структура 2000 года объясняет 99,08 % общей дисперсии и имеет следующий вид:

Значения главных компонент, 2000 год

№ компоненты	Доля предприятий отрасли	Объем производства в расчете на одного работника отрасли, тыс руб.	Среднее число работников на одном предприятии отрасли, чел.	Доля средней заработной платы по отрасли в процентах от средней по всем малым предприятиям	Процент объясняемой дисперсии
1	-0,094	0,991	0,003	0,989	49,08
2	0,990	-0,001	-0,136	-0,056	25,00
3	-0,136	-0,022	0,991	0,031	25,00

Интерпретация компонент совпадает с той, что получена для 1999 года. Рассмотрим факторную структуру 2001 года:

Значения главных компонент, 2001 год

№ компоненты	Доля предприятий отрасли	Объем производства в расчете на одного работника отрасли, тыс руб.	Среднее число работников на одном предприятии отрасли, чел.	Доля средней заработной платы по отрасли в процентах от средней по всем малым предприятиям	Процент объясняемой дисперсии
1	0,079	0,970	-0,227	0,971	48,08
2	-0,017	-0,138	0,974	-0,252	25,78
3	0,997	0,175	-0,019	-0,029	25,63

Структура компонент 2001 года также совпадает с выделенными в 1999–2000 годах. Однако вторая и третья компоненты поменялись местами, вследствие изменившегося соотношения в объясняемой дисперсии, но эти изменения весьма незначительны.

Таким образом, проведенный компонентный анализ позволил выявить внутреннюю взаимосвязь переменных, характеризующих отраслевую структуру малых предприятий. Первая и сходная по всем годам компонента, связывая отраслевой объем производства товаров и услуг, пересчитанный на одного работника отрасли со средней по отрасли заработной платой, характеризует относительную отраслевую доходность. Следующая компонента имеет значительную факторную нагрузку только при переменной числа малых предприятий в отрасли и интерпретирована как показатель масштаба. Численность работников в расчете на одно предприятие отрасли не связана с другими компонентами и отражает специфику производства в отрасли.

Поскольку выделенная первая компонента во всех изучаемых периодах объясняла основную долю дисперсии, то можно предположить, что именно этот фактор достаточно устойчиво характеризует относительную доходность функционирования малых предприятий в отраслях экономики.

Пример 6.4. Для условий примера 5.2 выполнить компонентный анализ и затем классифицировать наблюдения по главным компонентам на 5 групп.

Решение. Воспользуемся пакетом прикладных программ Statistica. После нормирования данных получим следующие результаты. Первые две главные компоненты объясняют около 81% общей дисперсии. Их вычисленные значения приведены в таблице:

X_i	Y_1	Y_2
age	-0,68	0,46
edu	0,06	-0,96
nhh	0,91	0,01
% объясняемой дисперсии	53	28

Первая главная компонента интерпретируется как число членов домохозяйства, вторая — как образование главы домохозяйства.

Далее воспользовавшись значениями первых двух главных компонент по домохозяйствам и методом k-средних, получим классификацию на 5 кластеров. Матрица евклидовых расстояний (под диагональю) и квадратов расстояний (над диагональю) между кластерами: $D = \begin{bmatrix} 0 & 3,14 & 3,08 & 4,27 & 1,08 \\ 1,77 & 0 & 0,78 & 3,79 & 2,47 \\ 1,75 & 0,89 & 0 & 1,15 & 1,09 \\ 2,07 & 1,95 & 1,07 & 0 & 1,06 \\ 1,04 & 1,57 & 1,04 & 1,03 & 0 \end{bmatrix}$.

Дисперсионный анализ по переменным показывает, что две используемые для классификации компоненты значимы на 1% уровне.

Сравнение с результатами классификации в примере 5.2 показывает их сходство. Однако в данном примере мы использовали для анализа меньше переменных с несколько лучшими дискриминантными свойствами.

Построение регрессии логарифма заработной платы lw (предварительно нормируем эту переменную) на главные компоненты приводит к следующему результату (в круглых скобках — стандартные ошибки переменных): $lw = \underset{(0,014)}{-0,065Y_1} + \underset{(0,014)}{0,092Y_2}$, который показывает,

что логарифм заработной платы повышается при увеличении фактора Y_2 (домохозяйства с высоким уровнем образования главы) и понижается с ростом количества членов домохозяйства — Y_1.

Вопросы и задачи

1. Какие задачи решаются с помощью компонентного анализа?

2. Как находятся главные компоненты?

3. Как интерпретируются результаты компонентного анализа?

4. Деятельность предприятий региона характеризуется четырьмя показателями. При проведении компонентного анализа получены собственные числа: 1,2; 0,8; 0,4 и одно из них оказалось пропущенным. Чему оно равно? Чему равен относительный вклад двух первых компонент?

5. По данным о 5 домохозяйствах провести компонентный анализ на основе показателей удельного веса доходов, не связанных с основной работой, в общей сумме доходов x_1 и удельного веса расходов на питание x_2

№ п.п.	x_1	x_2
1	0,23	0,40
2	0,24	0,26
3	0,19	0,40
4	0,17	0,50
5	0,23	0,40

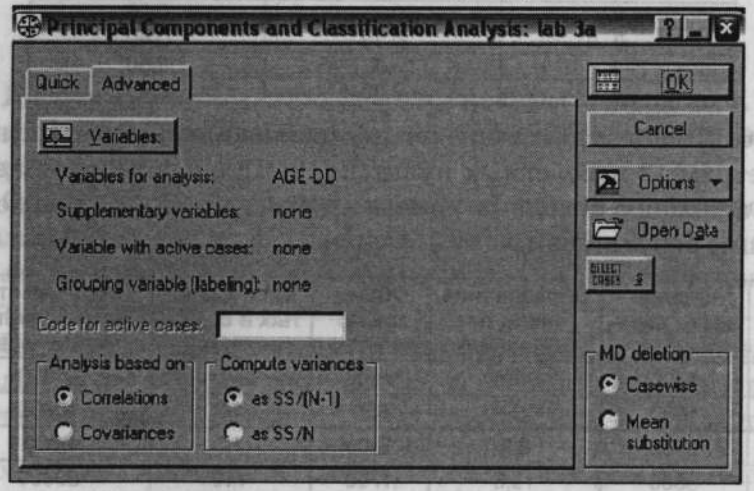
6. Имеются данные по пяти социально-экономическим параметрам 12 населенных пунктов [12]. Проведите компонентный анализ по исходным данным, дайте интерпретацию полученным результатам.

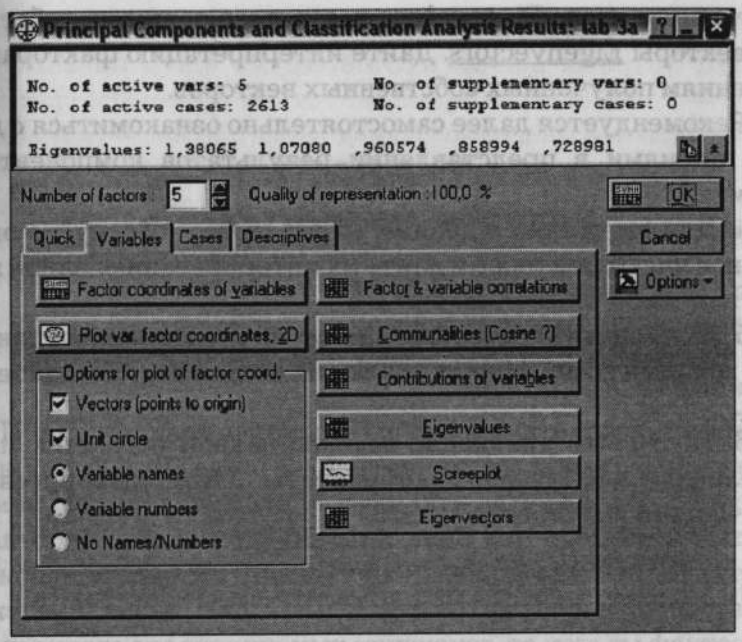
№ п.п.	Численность населения	Образование (число лет обучения)	Общее число занятых	Число занятых в сфере услуг	Средняя стоимость жилья (долларов)
1	5700	12,8	2500	270	25000
2	1000	10,9	600	10	10000
3	3400	8,8	1000	10	9000
4	3800	13,6	1700	140	25000

№ п.п.	Численность населения	Образование (число лет обучения)	Общее число занятых	Число занятых в сфере услуг	Средняя стоимость жилья (долларов)
5	4000	12,8	1600	140	25000
6	8200	8,3	2600	60	12000
7	1200	11,4	400	10	16000
8	9100	11,5	3300	60	14000
9	9900	12,5	3400	180	18000
10	9600	13,7	3600	390	25000
11	9600	9,6	3300	80	12000
12	9400	11,4	4000	100	13000

Задание к лабораторному практикуму
Время выполнения — 2 часа

1. Запустите приложение *Statistica*. Откройте файл *Lab3.sta*, воспользовавшись меню File\Open. Файл содержит подвыборку из массива msm.sta по индивидам. Описание переменных: *lw* — логарифм заработной платы, *edu* — число лет образования, *nhh* — число членов домохозяйства, *dd* — доля доходов в семейном бюджете, *age* — возраст, *pm* — процент заработков, который дает основная работа. Необходимо провести компонентный анализ данных, а затем классифицировать наблюдения.





2. Выполните расчет описательных статистик по переменным выборки. Сделайте выводы. Рассчитайте корреляционную матрицу переменных. Необходимо ли стандартизировать переменные?

3. Анализ главных компонент выполняется в программе Statistica с помощью модуля Statistics\Multivariate Exploratory Techniques\Principal Components & Classification Analysis. В появившемся окне Principal Components and Classification Analysis укажите переменные классификации и нажмите OK.

В верхней части появившегося окна результатов компонентного анализа дана общая информация и рассчитанные собственные числа корреляционной матрицы переменных.

Выберите число факторов (Number of factors) и получите соответствующий процент объясненной этими факторами вариации (Quality of representation). Просмотрите коэффициенты факторных нагрузок Factor&variable correlations. Далее можно получить график Screeplot собственных чисел и сами

собственные числа Eigenvalues, соответствующие им собственные векторы Eigenvectors. Дайте интерпретацию факторам по значениям полученных собственных векторов.

Рекомендуется далее самостоятельно ознакомиться с другими опциями в представлении результатов компонентного анализа.

4. Для имеющихся данных определите и обоснуйте оптимальное число факторов и дайте их интерпретацию. Какой процент вариации они объясняют?

5. Выполните кластерный анализ по значениям выбранных в п. 4 главных компонент для выборки индивидов. Сделайте выводы.

6. Постройте уравнение множественной регрессии *lw* на выделенные в п. 4 главные компоненты. Дайте интерпретацию полученного результата.

ГЛАВА 7
Факторный анализ

Вред или польза действия обуславливается совокупностью обстоятельств.

К. Прутков. Мысли и афоризмы.

Под факторным анализом будем понимать совокупность методов, которые на основе реально существующих связей признаков или объектов позволяют выявлять латентные обобщающие характеристики исследуемых явлений или процессов. Латентность характеристик означает их ненаблюдаемость, скрытость. Поскольку число общих (латентных) факторов существенно меньше числа анализируемых признаков, то методы факторного анализа в конечном счете нацелены (так же, как и метод главных компонент) на снижение размерности анализируемого признакового пространства. В зависимости от того, какой тип корреляционной связи исследуется, различают *R*, *Q*, *O*, *P*, *S* и *T* анализ. Чаще всего методами факторного анализа исследуются матрицы данных, в которых изучается вариация признаков (столбцы матрицы (1.1)) от объекта к объекту (строки) — *R-анализ*. Выявленные факторы интерпретируются как обобщенные характеристики, конденсирующие исходный набор признаков. Другой задачей является выявление группы объектов, имеющих сходный профиль по большому набору признаков, т.е. матрица (1.1) транспонируется и факторы интерпретируются как группы близких объектов и исследуется структура вариации объектов — это *Q-анализ*. Если рассматриваются наблюдения по признакам и во времени, то можно анализировать характер вариации признаков во времени и отыскивать факторы как некие обобщенные параметры, хорошо описывающие эту вариа-

цию. Такое представление матрицы (1.1) называют *P-анализом*. Столбцы матрицы данных соответствуют здесь признакам, а строки — временным интервалам. Тот же временной срез матрицы данных в транспонированном виде (столбцы — временные интервалы, строки — признаки) может обрабатываться методами факторного анализа с целью выявления периодов времени со сходным (внутри периода) сочетанием значений признаков. Такой подход носит название *O-анализа*. Существуют варианты факторного анализа, в которых рассматривается только один признак: в матрице исходных данных наблюдения по объектам и во времени (*S-анализ*). Такая постановка позволяет разрабатывать динамическую типологию объектов наблюдения, т.е. выявлять группы объектов со сходным типом изменений во времени. Транспонированная матрица (*T-анализ*) позволяет выявить периоды времени с характерным для каждого периода распределением значений исследуемого признака по объектам наблюдения.

В факторном анализе требуется, чтобы переменные измерялись в интервальной шкале, поскольку представление переменных в виде линейных комбинаций скрытых факторов для порядковых переменных невозможно. Если используется какой-либо способ шкалирования для порядковых переменных, не нарушающий их внутренних свойств (искажения матрицы корреляций не слишком велики), то можно использовать эти переменные в качестве числовых.

Дихотомические переменные нельзя представить в рамках факторной модели, поскольку в рамках последней каждая переменная является взвешенной суммой по крайней мере двух скрытых факторов (одного общего и одного специфического), следовательно, даже при двух значениях этих факторов (что не соответствует практике) наблюдаемая переменная будет принимать четыре возможных значения. Некоторые методы, применимые в этой ситуации, обсуждаются в [11. С. 64–65].

Рассмотрим основную модель факторного анализа, предполагая, что исходные данные предварительно стандартизованы вычитанием среднего и нормированием на стандартное отклонение.

7.1. Модель ортогональных факторов

Пусть мы наблюдаем вектор X размерности p со средними μ и ковариационной матрицей Σ.

Предположим, что X линейно зависит от нескольких ненаблюдаемых случайных переменных $F_1, F_2, \ldots, F_m$, называемых *общими факторами*, и дополнительно зависит от p источников вариации $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$, называемых специфическими факторами, или ошибками, которые не могут быть выражены через общие факторы, т.е.

$$\left.\begin{aligned} X_1-\mu_1 &= l_{11}F_1+l_{12}F_2+\ldots+l_{1m}F_m+\varepsilon_1 \\ X_2-\mu_2 &= l_{21}F_1+l_{22}F_2+\ldots+l_{2m}F_m+\varepsilon_2 \\ &\ldots \\ X_p-\mu_p &= l_{p1}F_1+l_{p2}F_2+\ldots+l_{pm}F_m+\varepsilon_p \end{aligned}\right\}, \tag{7.1}$$

или в матричном виде (7.1) имеет вид

$$\underset{p\times 1}{\mathbf{X}-\boldsymbol{\mu}} = \underset{p\times m}{\mathbf{L}}\ \underset{m\times 1}{\mathbf{F}} + \underset{p\times 1}{\boldsymbol{\varepsilon}}. \tag{7.2}$$

В (7.2) l_{ij} называется *нагрузкой* i-й переменой по j-му фактору и матрица L называется матрицей *факторных нагрузок*.

В (7.2) факторы F являются неизвестными случайными величинами. На практике число переменных p должно быть в два раза больше, чем факторов — m.

В (7.2) примем некоторые дополнительные ограничения на ε и F:

$$E(\mathbf{F}) = \underset{m\times 1}{\mathbf{0}}; \tag{7.3}$$

$$E(\mathbf{F}^T\mathbf{F}) = \underset{m\times m}{\mathbf{I}}; \tag{7.4}$$

$$E(\boldsymbol{\varepsilon}) = \underset{p\times 1}{\mathbf{0}}; \tag{7.5}$$

$$E(\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}) = \underset{p\times p}{\boldsymbol{\Psi}} = \begin{bmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \psi_p \end{bmatrix}. \tag{7.6}$$

$$E(\boldsymbol{\varepsilon}^T\mathbf{F}) = \underset{p\times m}{\mathbf{0}}. \tag{7.7}$$

Моделью с ортогональными факторами (заметим, что в этом случае они попарно некоррелированы) называется уравнение (7.2), при выполнении ограничений (7.3)-(7.7).

Найдем матричное произведение

$$(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T=(\mathbf{LF}+\boldsymbol{\varepsilon})(\mathbf{LF}+\boldsymbol{\varepsilon})^T=$$

$$=(\mathbf{LF}+\boldsymbol{\varepsilon})\left((\mathbf{LF})^T+\boldsymbol{\varepsilon}^T\right)=\mathbf{LF}(\mathbf{LF})^T+\boldsymbol{\varepsilon}(\mathbf{LF})^T+(\mathbf{LF})\boldsymbol{\varepsilon}^T+\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T. \quad (7.8)$$

Тогда: $\boldsymbol{\Sigma}=E\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right]=\mathbf{L}E\left(\mathbf{FF}^T\right)\mathbf{L}^T+$

$$+E\left(\boldsymbol{\varepsilon}\mathbf{F}^T\right)\mathbf{L}^T+\mathbf{L}E\left(\mathbf{F}\boldsymbol{\varepsilon}^T\right)+E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right)=\mathbf{LL}^T+\boldsymbol{\Psi}$$

$$\text{и}\quad \boldsymbol{\Sigma}=\mathbf{LL}^T+\boldsymbol{\Psi}. \quad (7.9)$$

Формула (7.9) задает ковариационную структуру модели ортогональных факторов и из нее, в частности, следует

$V(X_i)=l_{i1}^2+l_{i2}^2+\ldots+l_{im}^2+\psi_i$, $\operatorname{cov}(X_i,X_k)=l_{i1}l_{k1}+l_{i2}l_{k2}+\ldots+l_{im}l_{km}$,

$\operatorname{cov}(X_i,F_j)=l_{ij}$.

Модель факторного анализа (7.2) является линейной и содержательно означает, что наблюдаемые случайные величины представляются в виде линейной комбинации ненаблюдаемых общих факторов (F) и случайных ошибок ($\boldsymbol{\varepsilon}$), возникающих при таком представлении.

Задача заключается в нахождении матрицы факторных нагрузок L и содержательной интерпретации факторов $F_1,\ldots,F_m$.

Доля дисперсии i-й переменной, обусловленная m общими факторами, называются *общностью* h_i. Доля дисперсии ψ_i соответствующая специфическим факторам называется *специфической дисперсией*. Таким образом

$$\sigma_{ii}^2=l_{i1}^2+l_{i2}^2+\ldots+l_{im}^2+\psi_i=h_i^2+\psi_i, i=1,\ldots,p. \quad (7.10)$$

К сожалению, для факторного анализа ковариационная матрица не всегда может быть представлена в виде (7.9), т.е. решение может не существовать или не быть единственным.

Пусть $m>1$ и пусть T — ортогональная матрица размерности $m\times m$, причем $\mathbf{T}^T\mathbf{T}=\mathbf{TT}^T=\mathbf{I}$, тогда из (7.2) следует $\mathbf{X}-\boldsymbol{\mu}=\mathbf{LF}+\boldsymbol{\varepsilon}=\mathbf{LTT}^T\mathbf{F}+\boldsymbol{\varepsilon}=\mathbf{L}^*\mathbf{F}^*+\boldsymbol{\varepsilon}$, где $\mathbf{L}^*=\mathbf{LT}$ и $\mathbf{F}^*=\mathbf{T}^T\mathbf{F}$. Найдем математическое ожидание

$$E(\mathbf{F}^*)=\mathbf{T}^TE(\mathbf{F})=\mathbf{0},\operatorname{cov}(\mathbf{F}^*)=\mathbf{T}^T\operatorname{cov}(\mathbf{F})\mathbf{T}=\mathbf{T}^T\mathbf{T}=\mathbf{I}.$$

Таким образом, имеем, что факторы $\mathbf{F}$ и $\mathbf{F}^*=\mathbf{T}^T\mathbf{F}$ имеют одинаковые свойства и образованы одной и той же матрицей $\boldsymbol{\Sigma}$. Однако они имеют разные факторные нагрузки: в первом случае матрица факторных нагрузок — L, а во втором — матрица L^*, так что

$$\boldsymbol{\Sigma}=\mathbf{LL}^T+\boldsymbol{\Psi}=\mathbf{LTT}^T\mathbf{L}^T+\boldsymbol{\Psi}=\mathbf{L}^*\mathbf{L}^{*T}+\boldsymbol{\Psi}. \quad (7.11)$$

Таким образом, матрица T соответствует повороту системы координат, образованной переменными X, что означает *вращение факторов* $\mathbf{F}$ в этой системе координат.

Общности, лежащие на главной диагонали, либо $\mathbf{LL}^T$ матрицы, либо $\mathbf{L}^*\mathbf{L}^{*T}$, не зависят от выбора матрицы $\mathbf{T}$.

Вращение факторов помогает содержательно проинтерпретировать полученные факторы.

Главной задачей факторного анализа является определение элементов матрицы факторных нагрузок.

7.2. Определение факторных нагрузок методом главных факторов

Пусть имеются наблюдения $X_1,\ldots,X_n$ в p-мерном признаковом пространстве $X_i=\left(X_{i1}\quad X_{i2}\quad\ldots\quad X_{ip}\right)^T$.

Мы можем получить оценку ковариационной матрицы $\hat{\boldsymbol{\Sigma}}=\mathbf{S}$ этих переменных. Если недиагональные элементы матрицы $\mathbf{S}$, т.е. соответствующие парные коэффициенты корреляции малы по величине, то факторный анализ неприменим в силу плохой обусловленности матрицы $\mathbf{S}$.

В (7.2) необходимо оценить матрицу факторных нагрузок $\mathbf{L}$. Наиболее популярные методы нахождения оценок матрицы нагрузок: метод главных факторов (*Г. Томсон*), метод максимального правдоподобия (*Д. Лоули*), групповой метод (*Л. Гуттман* и *П. Хорст*), метод минимальных остатков (*Г. Харман*), метод канонического факторного анализа (*К. Рао*) и др.

Важное отличие метода главных факторов *от регрессионного анализа* заключается в том, что значения факторов

F ненаблюдаемы, а от *метода главных компонент* в том, что исходные данные описываются моделью факторного анализа вида (7.2) с относительно *малым* числом m, т.е. вариация исходных признаков может быть объяснена не на 100% (как в методе главных компонент, в котором число факторов равно числу исходных признаков), а несколько меньше, с учетом существования их нераскрываемой характерности. В этом случае корреляционная матрица *редуцирована*. Другими словами, в задаче поиска главных компонент критерием оптимальности служит минимальность отличия ковариационной матрицы главных компонент от ковариационной матрицы исходных признаков. Задача факторного анализа — наиболее полное объяснение корреляции между исходными признаками, т.е. каждый фактор подбирается из того условия, чтобы после его исключения из всех наблюдаемых признаков коэффициенты корреляции между всеми парами признаков были бы минимальными. Таким образом, если остаточные дисперсии (элементы $\mathbf{\Psi}$) невелики, то методы главных компонент и главных факторов должны давать близкие результаты.

Рассмотрим метод главных факторов для нахождения факторных нагрузок в (7.2). Пусть матрица $\mathbf{\Sigma}$ имеет собственные числа λ_i и собственные векторы e_i (λ_i, e_i), причем $\lambda_1 > \lambda_2 > \lambda_3 > \ldots > \lambda_p \geq 0$. Тогда матрица $\mathbf{\Sigma}$ представима в виде:

$$\mathbf{\Sigma} = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^T + \ldots + \lambda_p \mathbf{e}_p \mathbf{e}_p^T = \left(\sqrt{\lambda_1}\mathbf{e}_1 \vdots \sqrt{\lambda_2}\mathbf{e}_2 \vdots \ldots \vdots \sqrt{\lambda_p}\mathbf{e}_p\right) \begin{pmatrix} \sqrt{\lambda_1}\mathbf{e}_1^T \\ \cdots \\ \sqrt{\lambda_2}\mathbf{e}_2^T \\ \cdots \\ \cdots \\ \sqrt{\lambda_p}\mathbf{e}_p^T \end{pmatrix}. \tag{7.12}$$

Если $m=p$, тогда $\psi_i = 0$ для всех i и матрица $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{0} = \mathbf{L}\mathbf{L}^T$, т.е. j-й столбец матрицы $\mathbf{L}$ будет равен $\sqrt{\lambda_j}\mathbf{e}_j$.

Предположим, что m собственных чисел относительно большие, а $(p-m)$ относительно маленькие и будем рассматривать приближение к (7.12) вида:

$$\mathbf{\Sigma} = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T + \ldots + \lambda_m \mathbf{e}_m \mathbf{e}_m^T = \left(\sqrt{\lambda_1}\mathbf{e}_1 \vdots \sqrt{\lambda_2}\mathbf{e}_2 \vdots \ldots \vdots \sqrt{\lambda_m}\mathbf{e}_m\right) \begin{pmatrix} \sqrt{\lambda_1}\mathbf{e}_1^T \\ \cdots \\ \sqrt{\lambda_2}\mathbf{e}_2^T \\ \cdots \\ \cdots \\ \sqrt{\lambda_m}\mathbf{e}_m^T \end{pmatrix} = \underset{p\times m}{\mathbf{L}}\ \underset{m\times p}{\mathbf{L}^T}. \tag{7.13}$$

В (7.13) специфические факторы предполагаются $\varepsilon=0$. Если они учитываются в модели, то значения их дисперсий получаются стоящими на диагонали матрицы $\mathbf{\Psi} = \mathbf{\Sigma} - \mathbf{L}\mathbf{L}^T$, т.е. модель имеет вид

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi} = \left(\sqrt{\lambda_1}\mathbf{e}_1 \vdots \sqrt{\lambda_2}\mathbf{e}_2 \vdots \ldots \vdots \sqrt{\lambda_m}\mathbf{e}_m\right) \begin{pmatrix} \sqrt{\lambda_1}\mathbf{e}_1^T \\ \cdots \\ \sqrt{\lambda_2}\mathbf{e}_2^T \\ \cdots \\ \cdots \\ \sqrt{\lambda_m}\mathbf{e}_m^T \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ \vdots & \ldots & \ddots & \vdots \\ 0 & 0 & \ldots & \psi_p \end{pmatrix}, \tag{7.14}$$

где $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2, i = 1, \ldots, p.$ (7.15)

Таким образом, элементы матрицы $\mathbf{L}$ (факторные нагрузки) определяются по алгоритму:

1) исходные данные стандартизируются и тогда ковариационная матрица переходит в корреляционную матрицу. Использование корреляционной матрицы оправдано в случаях, когда дисперсии переменных существенно отличаются одна от другой и имеют разнородные единицы измерения, поскольку факторные шкалы в этой ситуации будет трудно интерпретировать;

2) определяется *редуцированная* матрица $\mathbf{\rho}_h$, в которой на главной диагонали стоят значения $h_j^2 < 1$, т.е. вариация признаков $x_j, j = 1, \ldots, p$ может быть объяснена не на 100%, а несколько меньше с учетом существования специфических факторов ε. Имеется несколько простых методов поиска общностей h_j^2:

- метод наибольшей корреляции — в строке (или столбце) находится наибольший по абсолютной величине коэффициент корреляции и принимается в качестве h_j^2;

- при помощи квадрата коэффициента множественной корреляции — $h_j^2 = 1 - \frac{1}{r_{jj}}$, где r_{jj} — диагональный элемент обратной к ρ матрицы;
- метод оценки h_j^2 при помощи среднего коэффициента корреляции по строке (столбцу);
- метод триад — в j-й строке (столбце) отыскиваются два наибольших значения коэффициентов корреляции r_{jk} и r_{jl} и составляется триада: $h_j^2 = \frac{r_{jk} r_{jl}}{r_{kl}}$;
- метод малого центроида. Для переменной j строится корреляционная матрица 4×4, включающая саму эту переменную и еще три, наиболее тесно связанные с ней. По данным матрицы и рассчитываются общности $h_j^2 = \frac{\left(\sum_i r_{il}\right)^2}{\sum_{ij} r_{ij}}$;

3) находятся собственные числа и собственные векторы матрицы ρ_h;

4) из найденных собственных чисел оставляем первые m, объясняющие наибольшую долю вариации. Матрица факторных нагрузок получается по (7.13): $\mathbf{L} = \left(\sqrt{\lambda_1}\mathbf{e}_1 \vdots \sqrt{\lambda_2}\mathbf{e}_2 \vdots \ \dots \ \vdots \sqrt{\lambda_m}\mathbf{e}_m\right)$.

Оценка специфических факторов ψ_i выполняется по формуле (7.15), причем в этом случае общности h_i^2 находятся по формуле:

$$h_i^2 = l_{i1}^2 + l_{i2}^2 + \dots + l_{im}^2, i = 1,..,m. \qquad (7.16)$$

Алгоритмически и содержательно метод главных компонент является частным случаем более общей модели факторного анализа и метода главных факторов. В первом случае преобразования применяются к исходной матрице корреляций, во втором — к редуцированной.

В факторном анализе интерес представляют не только факторные нагрузки — элементы матрицы **L**, но и оценки значений общих факторов, которые часто используются для последующего анализа. Необходимо оценить значения $\hat{\mathbf{f}}_j$, $j=1,..,n$, полученные как величина i-го фактора для j-го наблюдения: $\hat{\mathbf{f}}_j = \left(F_{1j}, F_{2j}, \dots, F_{mj}\right)^T$, $j=1,..,n$.

Для нахождения значений факторов предполагается, что получены нагрузки факторов **L** и оценки специфических дисперсий Ψ.

Исходя из модели факторного анализа (7.2) и предполагая, что ψ_i неодинаковые, Бартлетт предложил следующую процедуру для нахождения значений факторов (взвешенным методом наименьших квадратов). Минимизируя сумму квадратов ошибок, взвешенную на обратную величину их дисперсии:

$$\sum_{i=1}^{p} \frac{\varepsilon_i^2}{\psi_i} = \boldsymbol{\varepsilon}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\varepsilon} = \left(\mathbf{X} - \boldsymbol{\mu} - \mathbf{L}\mathbf{f}\right)^T \boldsymbol{\Psi}^{-1} \left(\mathbf{X} - \boldsymbol{\mu} - \mathbf{L}\mathbf{f}\right) \qquad (7.17)$$

и используя в (7.17) оценки необходимых матриц, получим:

$$\hat{\mathbf{f}}_j = \left(\hat{\mathbf{L}}^T \hat{\boldsymbol{\Psi}}^{-1} \hat{\mathbf{L}}\right)^{-1} \hat{\mathbf{L}}^T \hat{\boldsymbol{\Psi}}^{-1} \left(\mathbf{X}_j - \hat{\boldsymbol{\mu}}\right). \qquad (7.18)$$

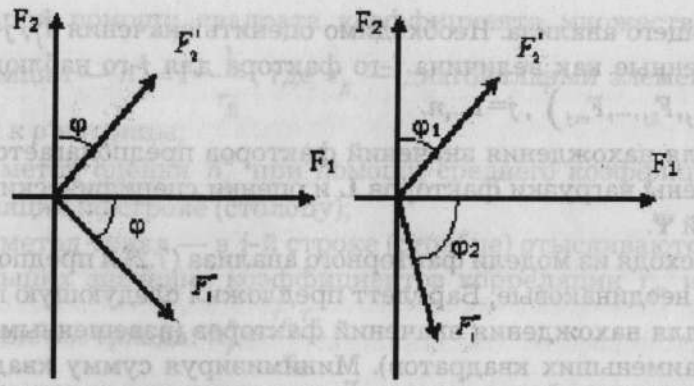
Отметим, что в случае получения факторных нагрузок методом главных факторов, обычно применяется метод наименьших квадратов (невзвешенный), в предположении, что ψ_i почти одинаковы: $\hat{\mathbf{f}}_j = \left(\hat{\mathbf{L}}^T \hat{\mathbf{L}}\right)^{-1} \hat{\mathbf{L}}^T \left(\mathbf{X}_j - \hat{\boldsymbol{\mu}}\right)$.

Отметим также, что сам по себе знак факторных нагрузок не несет информации о зависимости между переменной и фактором. Однако следует сопоставлять между собой знаки для различных переменных при данном факторе.

7.3. Вращение пространства общих факторов

Задача вращения общих факторов осуществляется с целью улучшения их интерпретируемости.

Из (7.11) следует, что можно от исходной матрицы нагрузок факторов перейти к другой матрице нагрузок $\mathbf{L}^*$, при этом значения общностей не меняются. Таким образом, мы можем изменять начальные нагрузки факторов путем поворота координатных осей, образованных этими факторами.



а) б)

Рис. 7.1. *Вращение общих факторов*

Если факторы ортогональны друг другу и вращение осуществляется на один и тот же угол для всех осей, то имеем ортогональное вращение (рис. 7.1, а). Если вращение осей осуществляется на произвольные углы, то имеем косоугольное вращение (рис. 7.1, б).

Вращение факторов необходимо для лучшей интерпретации полученных факторов, а именно хотелось бы в идеале, чтобы наблюдения располагались в пространстве факторов, как можно ближе к какой-либо из осей (фактору).

Чаще всего применяется ортогональное вращение факторов (рис. 7.1, а). При m=2 могут быть применены для вращения следующие матрицы:

$$T = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}$$ — вращение по часовой стрелке,

$$T = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$$ — вращение против часовой стрелки.

Если $m > 2$, графическая визуализация затруднена и используются матрицы T, имеющие похожую структуру. Выбор матрицы T и достаточность числа поворотов пространства осуществляются на основе критериев для оценки структуры общих факторов. Наиболее популярны критерии:

квартимакс: $q = \sum_{j=1}^{p} \frac{\sum_{k=1}^{m} l_{jk}^4 - \left(\sum_{k=1}^{m} l_{jk}^2\right)^2}{m}$, факторные нагрузки должны максимизировать q;

варимакс: $V_m = \frac{m\sum_{k=1}^{m} l_{jk}^4 - \left(\sum_{k=1}^{m} l_{jk}^2\right)^2}{m^2}$, который рассчитывается для каждого фактора и позволяет достичь упрощения структуры в описании столбцов матрицы **L**, наилучшим считается максимальное значение критерия;

облимакс: $K = \sum_{j=1}^{p}\sum_{k=1}^{m} l_{jk}^4 \Big/ \left(\sum_{j=1}^{p}\sum_{k=1}^{m} l_{jk}^2\right)^2$, максимизируя K, оценивается улучшение элементов структуры **L** каждого общего фактора и по всем факторам.

Как следует из п 7.1-7.2, собственные значения, связанные с факторами до вращения, не совпадают с соответствующими величинами для вращаемых факторов — неизменна только сумма собственных значений. В первоначальном факторном решении величина собственного значения несет информацию об относительной важности каждого фактора. Для факторного решения после вращения это свойство не сохраняется.

7.4. Статистическая оценка надежности решений методом факторного анализа

Корректное решение задач при помощи методов факторного анализа предполагает подтверждение значимости исходной матрицы парных корреляций (ковариаций) и достаточности числа обобщенных факторных признаков в анализе.

Значимость корреляционной матрицы подвергается проверке, если принять во внимание, что незначимые корреляционные (ковариационные) связи элементарных признаков

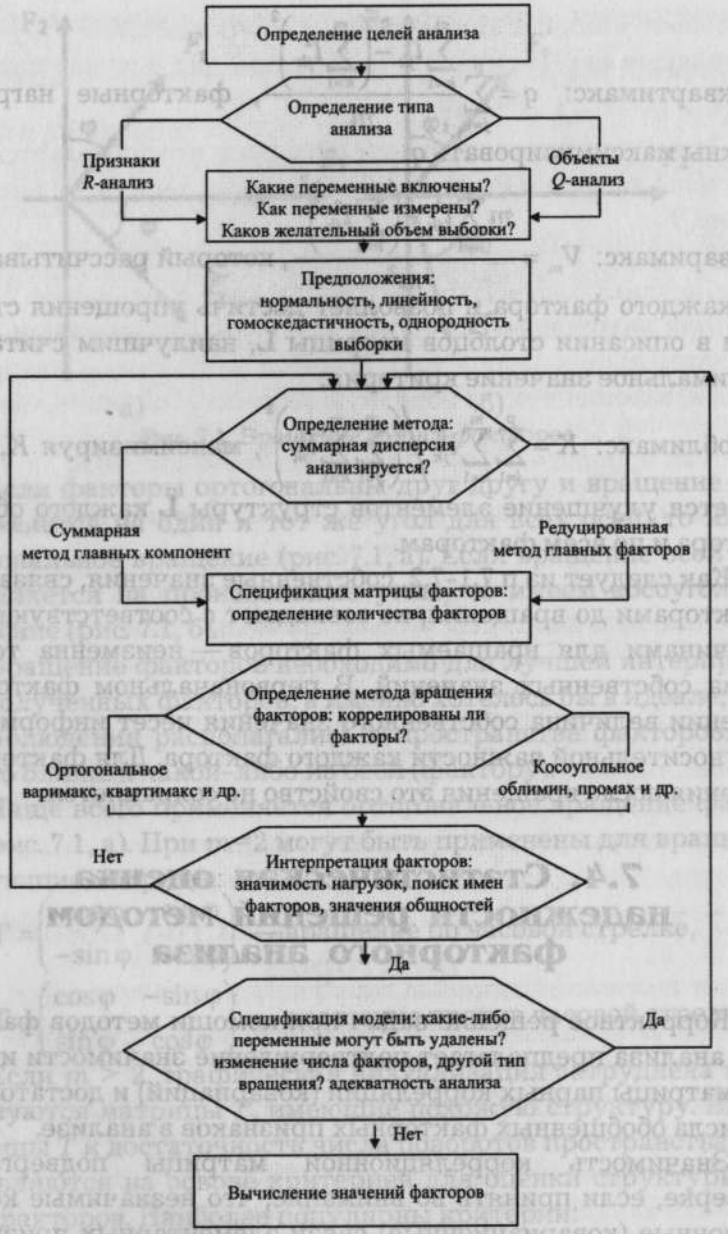


Рис. 7.2. *Блок-схема принятия решений в факторном анализе*

не дают оснований для поиска обобщенных признаков. В этом случае все вычисленные собственные числа будут близки к единице, и обобщенные факторы по составу становятся очень похожими на элементарные признаки. Проведение факторного анализа теряет смысл (см. рис. 7.2.).

Проверка значимости матрицы парных корреляций осуществляется при помощи критерия Уилкса, наблюдаемое значение которого находится по формуле: $\chi_s^2 = -\left(n - \frac{1}{6}(2p+5)\right)\ln(\det(\boldsymbol{\rho}))$, где $\boldsymbol{\rho}$ — матрица парных корреляций; n, p — соответственно число наблюдаемых объектов и число элементарных признаков в анализе.

Определитель $\det(\boldsymbol{\rho}) = \prod_{j=1}^{p} \lambda_j$. Если табличное значение χ^2 распределения при заданном уровне значимости и числе степеней свободы $0{,}5p(p-1)$ меньше расчетного, то значимость корреляционной матрицы подтверждается.

Для оценки достаточности числа выделенных общих признаков (факторов) в методе главных компонент используется критерий Бартлетта: $\chi_s^2 = -\left(n - \frac{1}{6}(2p+5) - \frac{2}{3}k\right)\ln \rho_{p-k}$, здесь k — число оставленных в анализе главных компонент;

$$\rho_{p-k} = \frac{\det(\boldsymbol{\rho})}{\prod_{j=1}^{k} \lambda_k \left(\frac{p - \sum_{j=1}^{k} \lambda_k}{p-k}\right)^{p-k}}.$$

Число степеней свободы для критерия χ^2 будет $0{,}5((p-k)-p-k-1)$ и если наблюдаемое значение меньше табличного, то принимается предположение о том, что выделенные k главных компонент достаточно полно представляют дисперсию p элементарных признаков и остальные $p-k$ компонент могут в анализе не рассматриваться из-за незначительного уровня их информативности. Иначе, в анализ для формирования выводов должны быть введены дополнительно другие главные компоненты.

В факторном анализе для проверки гипотезы о достаточности числа обобщенных признаков (факторов) используется χ^2-критерий Лоули, имеющий аналогичную, как и в предыдущем случае, смысловую нагрузку: $\chi_s^2 = (n-1)\ln\dfrac{\det(\mathbf{LL}^T)}{\det(\boldsymbol{\rho})}$.

Критическое значение при заданном уровне значимости и числе степеней свободы $0{,}5((p-m)^2-p-m)$. Предположение о достаточном числе общих факторов подтверждается, когда $\chi_s^2 < \chi_\varepsilon^2$. Эмпирически Лоули получено, что χ^2-аппроксимация точнее, когда выборка содержит на 51 наблюдение больше, чем число переменных.

Для оценки адекватности факторной модели в целом может использоваться подход Хармана. К сожалению, он не содержит рекомендаций по поводу пороговых уровней, скажем неадекватности, но в сравнении с другими приемами значительно легче при реализации и базируется на простой средней оценке расхождений исходных и воспроизведенных коэффициентов корреляции: $\sum\sum_{j\neq i}\left(r_{ij}-r_{ij}^{+}\right)^2 \Big/ p(p-1)$.

Средний квадрат отклонений Хармана исчисляется по всем, кроме диагональных, коэффициентам корреляции. Из нескольких моделей факторного анализа, естественно, будет лучше та, для которой средняя сумма квадратов отклонений окажется наименьшей.

Пример 7.1. Пусть дана матрица $\Sigma = \begin{bmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{bmatrix}$.

Тогда факторное разложение

$$\begin{bmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} 4 & 7 & -1 & 1 \\ 1 & 2 & 6 & 8 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} = \mathbf{LL}^T + \boldsymbol{\Psi}.$$

Общность, например, для X_1: $l_{11}^2 + l_{12}^2 = 4^2 + 1^2 = 17$. Дисперсия, например, $\sigma_{11}^2 = h_1^2 + \psi_1 = 17 + 2 = 19$.

Если $m=p$, то матрица $\boldsymbol{\Sigma} = \mathbf{LL}^T + \boldsymbol{\Psi} = \mathbf{LL}^T$ и $\boldsymbol{\Psi} = 0$.

Предположим, что $m=2$, $p=12$, тогда $\dfrac{p(p+1)}{2} = 78$ элементов матрицы $\boldsymbol{\Sigma}$ описывается ($mp+p=36$) 36-ю параметрами: l_{ij} и ψ_i факторной модели.

Пример 7.2. Для условий примера 7.2 выполнить факторный анализ, предполагая $m=2$.

Решение.

Полученную корреляционную матрицу

$\boldsymbol{\rho} = \begin{bmatrix} 1 & -0{,}125 & -0{,}090 \\ -0{,}125 & 1 & -0{,}593 \\ -0{,}090 & -0{,}593 & 1 \end{bmatrix}$ заменяем редуцированной, воспользовавшись методом наибольшего коэффициента корреляции: $\boldsymbol{\rho}_h = \begin{bmatrix} 0{,}125 & -0{,}125 & -0{,}090 \\ -0{,}125 & 0{,}593 & -0{,}593 \\ -0{,}090 & -0{,}593 & 0{,}593 \end{bmatrix}$. Находим первые два собственных числа матрицы $\lambda_1 = 1{,}186$, $\lambda_2 = 0{,}226$ и собственные векторы $\mathbf{e}_1 = (0{,}023\ -0{,}709\ 0{,}705)$, $\mathbf{e}_2 = (0{,}830\ -0{,}380\ -0{,}409)$. Тогда матрица факторных нагрузок с учетом (7.13) будет иметь вид:

$\mathbf{L} = \begin{bmatrix} 0{,}026 & 0{,}395 \\ -0{,}772 & -0{,}180 \\ 0{,}768 & -0{,}195 \end{bmatrix}$. Общности $h_1^2 = 0{,}157$, $h_2^2 = 0{,}628$, $h_3^2 = 0{,}628$.

Пример 7.3. Исходные данные из массива msm.sta. Файл содержит подвыборку из 2529 наблюдений. Описание переменных (для главы домохозяйства): *nhh* — число членов домохозяйства, *dd* — доля доходов главы домохозяйства в семейном бюджете, *pm* — процент заработков, который дает основная работа, *pt* — процент времени, который занимает основная работа, e_1 — количество изменений основного места работы, e_2 — количество изменений основной профессии, e_3 — количество изменений должности. Необходимо выполнить факторный анализ, предполагая $m=2$.

Решение.

Воспользуемся пакетом прикладных программ Statistica. После нормирования данных получим следующие результаты. На рисунке представлены собственные числа, полученные методом максимального правдоподобия, для двух факторов

Eigenvalues (example 7_3)
Extraction: Maximum likelihood factors

Value	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	1,937299	27,67570	1,937299	27,67570
2	1,281242	18,30345	3,218541	45,97916

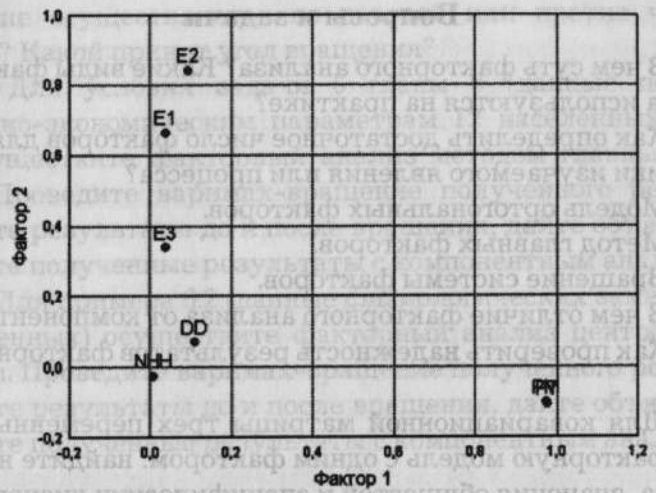
а также коэффициенты матрицы факторных нагрузок,

Factor Loadings (Varimax normalized) (example 7_3)
Extraction: Maximum likelihood factors
(Marked loadings are > ,700000)

Variable	Factor 1	Factor 2
PT	0,977638	-0,106018
PM	0,973956	-0,100425
E1	0,037578	0,663902
E2	0,093152	0,841194
E3	0,036495	0,339455
NHH	0,008168	-0,025110
DD	0,110134	0,070711
Expl.Var	1,927983	1,290558
Prp.Totl	0,275426	0,184365

которые поддаются интерпретации как фактор 1 — важность основной работы по времени занятости и доходам, фактор 2 — количество изменений профессии индивида.

Расположение переменных в пространстве полученных факторов показано на рисунке (после метода вращения варимакс):



По рисунку видно, что переменные *pt*, *pm* и *e*2, участвовавшие в формировании факторов, формируют оси системы координат, в которой располагаются остальные переменные.

Корреляционная матрица для остатков Ψ показывает, что имеется один значимый внедиагональный коэффициент корреляции, что свидетельствует о неплохом качестве проведенного факторного анализа.

Residual Correlations (example 7_3)
Extraction: Maximum likelihood factors
(Marked residuals are > ,100000)

Variable	PT	PM	E1	E2	E3	NHH	DD
PT	0,03	0,00	-0,00	-0,00	0,00	0,00	0,01
PM	0,00	0,04	0,00	0,00	-0,01	-0,00	-0,01
E1	-0,00	0,00	0,56	0,00	-0,01	0,00	0,00
E2	-0,00	0,00	0,00	0,28	0,00	0,01	-0,01
E3	0,00	-0,01	-0,01	0,00	0,88	-0,00	0,05
NHH	0,00	-0,00	0,00	0,01	-0,00	1,00	-0,23
DD	0,01	-0,01	0,00	-0,01	0,05	-0,23	0,98

Также могут быть получены значения факторов для каждого наблюдения, которые можно использовать, в частности, при проведении кластерного анализа домохозяйств по выделенным факторам.

Вопросы и задачи

1. В чем суть факторного анализа? Какие виды факторного анализа используются на практике?

2. Как определить достаточное число факторов для характеристики изучаемого явления или процесса?

3. Модель ортогональных факторов.

4. Метод главных факторов.

5. Вращение системы факторов.

6. В чем отличие факторного анализа от компонентного?

7. Как проверить надежность результатов факторного анализа?

8. Для ковариационной матрицы трех переменных постройте факторную модель с одним фактором: найдите нагрузки фактора, значения общностей и специфических дисперсий:

$$\Sigma = \begin{bmatrix} 1 & 0{,}4 & 0{,}9 \\ 0{,}4 & 1 & 0{,}7 \\ 0{,}9 & 0{,}7 & 1 \end{bmatrix}.$$

9. По известной матрице факторных нагрузок $\begin{bmatrix} 0{,}76 & 0{,}42 \\ 0{,}45 & 0{,}21 \\ 0{,}65 & 0{,}37 \\ 0{,}38 & 0{,}19 \end{bmatrix}$ воспроизведите матрицу парных корреляций.

10. Имеются матрица факторных нагрузок

$$\mathbf{L} = \begin{bmatrix} 0{,}86 & -0{,}25 \\ 0{,}61 & -0{,}31 \\ 0{,}44 & 0{,}51 \\ -0{,}47 & -0{,}28 \\ 0{,}38 & 0{,}08 \end{bmatrix}$$ и матрица вращения $\mathbf{T} = \begin{bmatrix} 0{,}574 & 0{,}819 \\ -0{,}819 & 0{,}574 \end{bmatrix}$.

Матрица $\boldsymbol{T}$ — ортогонального или косоугольного вращения? Вращение осуществляется по часовой или против часовой стрелке? Какой принят угол вращения?

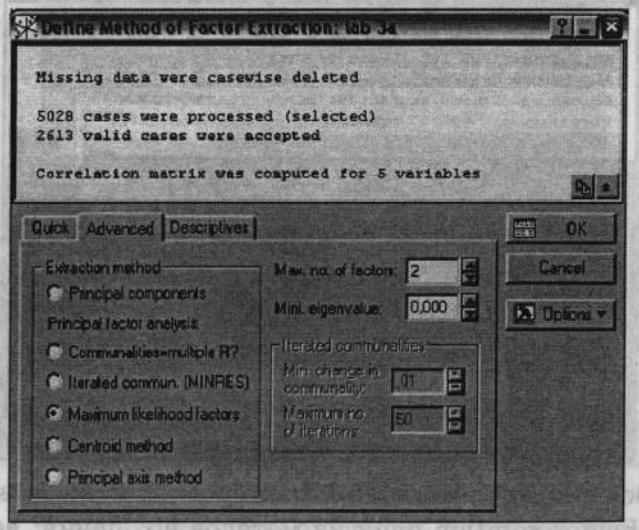
11. Для условия задачи 6 главы 6 (данные по пяти социально-экономическим параметрам 12 населенных пунктов) осуществите факторный анализ методом главных факторов. Проведите варимах-вращение полученного решения. Сравните результаты до и после вращения, дайте объяснение. Сравните полученные результаты с компонентным анализом.

12. Для примера 6.2 (данные физиологических замеров новорожденных) осуществите факторный анализ центроидным методом. Проведите варимах-вращение полученного решения. Сравните результаты до и после вращения, дайте объяснение. Сравните полученные результаты с компонентным анализом

Задание к лабораторному практикуму
Время выполнения — 4 часа

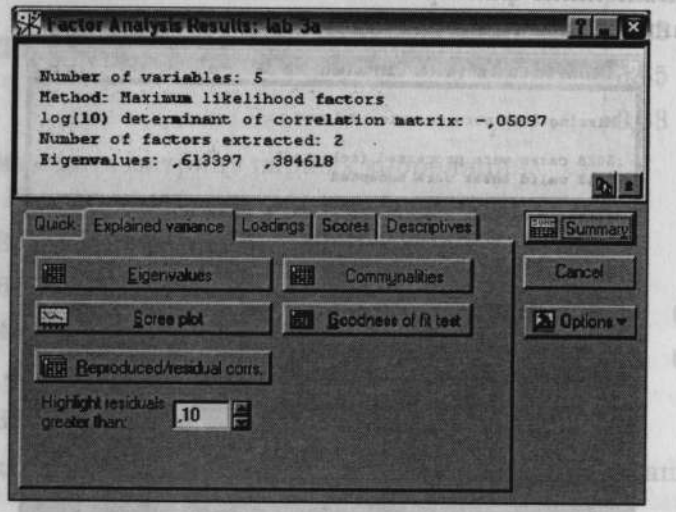
1. Рабочий файл данных тот же, что и в предыдущей лабораторной работе: *Lab3.sta*.

2. Выполним факторный анализ по имеющимся выборочным данным.



В программе Statistica запустите модуль Statistics\Multivariate Exploratory Techniques\Factor Analysis и в появившемся окне укажите переменные для анализа (укажите все имеющиеся переменные, кроме *lw*) и нажмите OK. Далее во вкладке Advanced укажите метод поиска латентных факторов, например, Principal factor analysis\Maximum likelihood factors, метод главных факторов (рекомендуется сравнить результаты, полученные разными методами оценивания), а также максимальное количество факторов, например 2, и минимальное собственное число фактора для включения его в анализ — 0. Нажмите OK.

В появившейся таблице результатов после нажатия кнопки Summary на вкладке Quick получим факторные нагрузки, причем красным цветом выделены коэффициенты, большие по модулю 0,7. Также можно получить (опция Eigenvalues) значения собственных чисел и долю объясненной факторами вариации. Воспользовавшись вкладкой Explained variance и опцией Communalities, получите значения общностей для каждого из фактора. С помощью вкладки Scores и кнопки Factor scores получим значения факторов для каждого наблюдения.



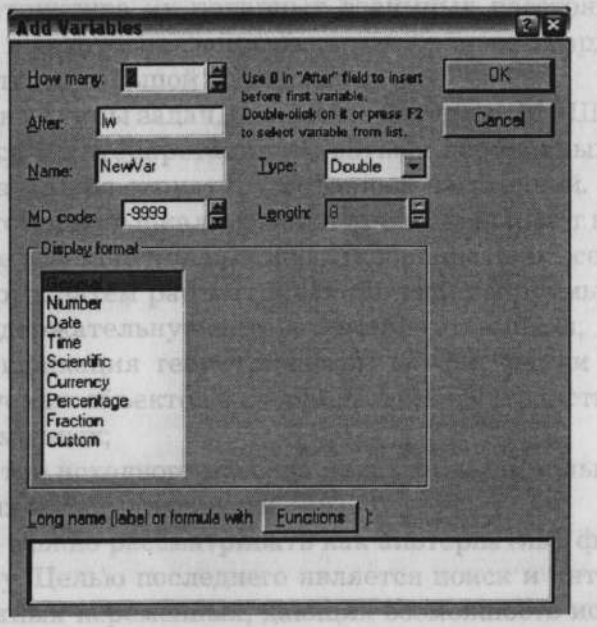
3. Осуществим вращение факторов. Для этого выберите в меню Factor rotation один из методов вращения: варимакс,

биквартимакс, квартимакс, эквимакс, например, Varimax normalized. Посмотрите, как изменились факторные нагрузки и другие результаты факторного анализа. Дайте интерпретацию факторов.

Поэкспериментируйте, выбирая различные значения числа факторов на начальной стадии анализа и различные методы вращения факторов. Выберите оптимальный с вашей точки зрения результат.

4. Выполните кластерный анализ наблюдений (по объектам) по факторам, полученным после варимакс-вращения в предыдущем пункте. Для этого выполните следующие действия.

В окне исходных данных, щелкнув правой кнопкой мыши по заголовку любой переменной, выберите Add variables и добавьте необходимое количество (по числу факторов в п. 3) новых переменных в конец списка переменных:



Вернувшись в окно факторного анализа с полученными как в пункте 3 факторами с помощью вкладки Scores и кнопки Factor scores, получите таблицу значений факторов для каждого

наблюдения. Скопируйте два столбца значений таблицы в окно исходных данных на место добавленных новых переменных (с именами по умолчанию NewVar — можно переименовать переменную, дважды кликнув по заголовку). Далее воспользуйтесь меню Statistics\Multivariate Exploratory Techniques\ Cluster Analysis и методом автоматической классификации k-средних (см. лабораторную работу по кластерному анализу) для вновь полученных переменных.

5. Постройте уравнение множественной регрессии lw на выделенные факторы.

6. Постройте необходимые графики и рассчитайте соответствующие статистики. Сделайте содержательные экономические выводы по результатам статистического анализа.

ГЛАВА 8
Многомерное шкалирование

Самый отдалённый пункт земного шара к чему-нибудь да близок, а самый близкий от чего-нибудь да отдалён.

К. Прутков. Мысли и афоризмы

Многомерное шкалирование (МНШ) — методы, реализующие «погружение» анализируемых объектов, о которых известна лишь структура их попарных взаимных расстояний (мер сходства), в некоторое специально подобранное координатное пространство небольшой размерности.

Основные типы задач, решаемые методами МНШ:

– поиск и интерпретация латентных переменных, объясняющих заданную структуру попарных расстояний. Этот тип задач многомерного шкалирования предусматривает не только построение вспомогательных шкал (координатных осей), в системе которых затем рассматриваются анализируемые объекты, но и содержательную интерпретацию этих шкал;

– верификация геометрической конфигурации системы анализируемых объектов в координатном пространстве латентных переменных;

– сжатие исходного массива данных с минимальными потерями в их информативности.

МНШ можно рассматривать как альтернативу факторному анализу. Целью последнего является поиск и интерпретация латентных переменных, дающих возможность исследователю объяснить сходства между объектами, заданными точками в исходном пространстве признаков. В факторном анализе сходства/различия между объектами выражаются с помощью

матрицы коэффициентов корреляций. В методе МНШ дополнительно к корреляционным матрицам, в качестве исходных данных можно использовать произвольный тип матрицы сходства объектов. На входе алгоритма МНШ — матрица, элементы которой содержат сведения о попарном сходстве/различиях анализируемых объектов. На выходе алгоритма МНШ получаются числовые значения координат, которые приписываются каждому объекту в некоторой новой системе координат (во «вспомогательных шкалах», связанных с латентными переменными, откуда и название МНШ), причем размерность нового пространства признаков существенно меньше размерности исходного.

Основной тип данных в многомерном шкалировании — меры близости между двумя объектами, измеряющие насколько эти объекты похожи. Наиболее часто используются коэффициенты корреляции и совместные вероятности.

Обозначим меру близости двух объектов i и j как δ_{ij}. Если эта мера такова, что самые большие ее значения соответствуют парам наиболее похожих объектов, то δ_{ij} — мера сходства, если ее наибольшие значения соответствуют парам наименее похожих объектов, то δ_{ij} — мера различия (далее по умолчанию предполагается, что δ_{ij} — мера различия).

Необходимо, чтобы исходные данные представляли матрицу расстояний между объектами γ_{ij}, т.е. выполнен набор аксиом расстояний:

1) $\gamma_{ij} \geq 0, \gamma_{ii} = 0, i,j = 1,..,n$;

2) $\gamma_{ij} = \gamma_{ji}, i,j = 1,..,n$;

3) $\gamma_{ij} + \gamma_{jk} \geq \gamma_{ik}, i,j,k = 1,..,n$.

При этом часто матрица, оценивающая различия объектов, строится по результатам опросов экспертов на основе шкал сравнения или в категоризованной форме. В этом случае свойство 1) выполнено, а свойства 2) и 3) оказываются нарушенными. Для удовлетворения свойства 2) полагают $\gamma_{ij} = \gamma_{ji} = \frac{1}{2}\left(\delta_{ij} + \delta_{ji}\right)$. Если для мер различия δ_{ij} выполнены все свойства, кроме 3),

то можно определить величину $c = \max\limits_{k,i,j}\left(\delta_{kj} - \delta_{ki} - \delta_{ij}\right)$ и затем

$$\gamma_{ij} = \begin{cases} 0, & i = j, \\ \delta_{ij} + c, & i \neq j. \end{cases}$$

Если каждый из объектов задан набором числовых характеристик $\upsilon_{i1},...,\upsilon_{iq}$ и они одинаково важны для формирования различий объектов, то расстояние может быть определено как

$$\gamma_{ij} = \sqrt{\sum_{s=1}^{q}\left(\upsilon_{is} - \upsilon_{js}\right)^2}, i,j = 1,...,n.$$

Пусть исходная информация об объектах задана в форме матрицы попарных сравнений γ_{ij}. Если заданы коэффициенты корреляции, то можно положить $\gamma_{ij} = \sqrt{1 - r_{ij}}$, $i,j = 1,...,n$.

В классической модели МНШ, предложенной У. Торгерсоном [4], предполагается, что по этой матрице можно построить точки $x_1,...,x_n$ в пространстве некоторого небольшого числа измерений p, причем так, чтобы каждый элемент матрицы γ_{ij} представлял собой евклидово расстояние $\gamma_{ij} = \sqrt{\sum_{r=1}^{p}\left(x_i^{(r)} - x_j^{(r)}\right)^2}$ между объектом i и объектом j, т.е. в матричном виде:

$$\underset{n\times n}{\boldsymbol{\Gamma}} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} \end{bmatrix}. \qquad (8.1)$$

Неизвестны координаты объектов:

$$X_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}), i = 1,...,n. \qquad (8.2)$$

Требуется на основании известных расстояний вида (8.1) восстановить неизвестную размерность p пространства признаков и приписать каждому объекту координаты (8.2) так, чтобы вычисленные евклидовы расстояния совпали с заданными в (8.1).

Таким образом, цель МНШ в том, чтобы преобразовать информацию о конфигурации исходных многомерных данных,

заданную матрицей расстояний в геометрическую конфигурацию точек в многомерном пространстве.

Модель предполагает, кроме того, что $\sum_{i=1}^{n} x_i^{(r)} = 0, r = 1,..., p.$

Введем в рассмотрение матрицу $\mathbf{X}_ц$ центрированных значений координат объектов: $\mathbf{X}_ц = \begin{bmatrix} x_1^{(1)} - \bar{x}_1 & x_1^{(2)} - \bar{x}_2 & \dots & \bar{x}_1^{(p)} - \bar{x}_p \\ x_2^{(1)} - \bar{x}_1 & x_2^{(2)} - \bar{x}_2 & \dots & \bar{x}_2^{(p)} - \bar{x}_p \\ \dots & \dots & \dots & \dots \\ x_n^{(1)} - \bar{x}_1 & x_n^{(2)} - \bar{x}_2 & \dots & \bar{x}_n^{(p)} - \bar{x}_p \end{bmatrix}$ и матрицу такую, что

$$b_{ij} = (X_i - \bar{X})^T (X_j - \bar{X}) = \sum_{r=1}^{p} (x_i^{(r)} - \bar{x}^{(r)})(x_j^{(r)} - \bar{x}^{(r)}). \quad (8.3)$$

Можно показать, что элементы матрицы **Г** и **В** связаны между собой соотношением:

$$b_{ij} = \frac{1}{2}\left(-\gamma_{ij}^2 + \frac{1}{n}\sum_{i=1}^{n}\gamma_{ij}^2 + \frac{1}{n}\sum_{j=1}^{n}\gamma_{ij}^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}^2 \right). \quad (8.4)$$

Соотношение (8.3) представляет собой *теорему У. Торгерсона.*

Матрица **В** обладает рядом следующих свойств:

– неотрицательно определена;

– ранг матрицы **В** равен размерности p искомого пространства;

– ненулевые собственные числа матрицы **В**, упорядоченные по убыванию, совпадают с собственными числами матрицы $\mathbf{S}^2 = \mathbf{X}_ц^T \mathbf{X}_ц$;

– обозначим l_r, $r=1,..,p$ собственный вектор, соответствующий собственному значению λ_r матрицы $\mathbf{S}^2$, тогда вектор Z_r значений r-й главной компоненты вектора X будет определяться по правилу $Z_r = X_ц \cdot l_r$. Причем если l_r^B — собственный вектор матрицы B, соответствующий тому же собственному числу λ_r, то

$$Z_r = \sqrt{\lambda_r} l_r^B, r = 1,..,p. \quad (8.5)$$

Таким образом, из свойств матрицы **В** следует, что находя собственные числа и собственные векторы этой матрицы, получим координатное приставление (8.5) анализируемых объектов в пространстве главных компонент искомого многомерного признака X, что и является решением задачи МНШ. Также из свойств следует, что элементы матрицы **В** могут быть получены как скалярные произведения векторов $b_{ij} = \sum_{r=1}^{p} Z_i^{(r)} Z_j^{(r)}$.

В неметрическом многомерном шкалировании предполагается, что различие между объектами измерено в рангах. Процедура неметрического многомерного шкалирования строится так, чтобы ранговый порядок попарных расстояний между точками минимально отличался от того, который задан.

Дж. Краскалом для определения меры соответствия предложен следующий критерий, названный им «*стресс*» [4]:

$$S_1 = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\gamma_{ij}^{\cdot} - \gamma_{ij}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}^2}}$$, где $\gamma_{ij}^{\cdot}$ — расстояние между объектами в p-мерном пространстве.

Также используется критерий «стресс, формула 2»:

$$S_2 = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\gamma_{ij}^{\cdot} - \gamma_{ij}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\gamma_{ij} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}\right)^2}}.$$

Л. Гуттман предложил третью меру несоответствия — коэффициент отчуждения. Для определения коэффициента отчуждения вычисляется коэффициент монотонности:

$$\mu = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}^{\cdot}\gamma_{ij}}{\sqrt{\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\gamma_{ij}^{\cdot}\right)^2\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}^2\right)}},$$ который является ранговой мерой согласованности, указывающей степень порядковой связи между исходными данными и оценками расстояний. Затем определяется коэффициент отчуждения: $\kappa = \sqrt{1-\mu^2}$.

Чем лучше подогнана модель к данным, тем меньше будет к.

С учетом введенных мер несоответствия алгоритм МНШ имеет вид*:

1. Полагаем номер итерации $k=0$, $\hat{x}_i^{(r,k)} = z_i^{(r)}$, $\gamma_{ij}^{(k)} = \gamma_{ij}$, $i,j=1,...,n$, $r=1,...,p$, текущее значение стресс-критерия $S=0$.

2. Нормирование: $\hat{d}_{ij}^{(k)} = \sqrt{\sum_{r=1}^{p}\left(\hat{x}_i^{(r,k)} - \hat{x}_j^{(r,k)}\right)^2}$; $Q = \sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\hat{d}_{ij}^{(k)}\right)^2}$;

$d_{ij}^{(k)} = \hat{d}_{ij}^{(k)} / Q$, $x_i^{(r,k)} = \hat{x}_i^{(r,k)} / Q$, $i,j=1,...,n$, ;

$$S(k) = \sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\gamma_{ij}^{(k)} - d_{ij}^{(k)}\right)^2}.$$

3. Если $k>0$ и $|S(k)-S|<\varepsilon$, где ε — требуемая точность вычислений, то конец алгоритма, иначе полагаем $S=S(k)$ и переходим к следующему шагу.

4. Упорядочивание. Этот шаг выполняется при $k=0$. Упорядочим пары (i,j) по возрастанию различий γ_{ij}. Если для некоторых пар $\gamma_{ij}^{(k)} = \gamma_{pt}^{(k)}$ и $d_{ij}^{(k)} < d_{cd}^{(k)}$, то пара (i,j) должна предшествовать (c,d). Если равны и различия и расстояния, то порядок пар произвольный. В результате получим таблицу из C_n^2 строк и трех столбцов, в первом из которых — обозначение пары, во втором — различия для пар объектов в порядке возрастания, в третьем — расстояния между объектами. Алгоритм работает пока расстояния не окажутся упорядоченными. Полагаем $l=1$.

* Адаптировано из: *Дронов С.В.* Многомерный статистический анализ: Учебное пособие. — Барнаул: Изд-во Алт. гос. ун-та, 2003. — 213 с.

5. Если соседние в таблице числа $d_{ij}^{(k)}$ имеют равные величины, объединяем соответствующие им пары индексов в блок. Если равных расстояний нет, каждую пару считаем самостоятельным блоком.

6. Подсчитываем число блоков. Пусть их $m(l)$ и s-му блоку соответствует расстояние $d(s)$, $s=1,...,m(l)$. Положим $s=1$.

7. Сравниваем s-й и $(s+1)$-й блоки. Если $d(s)<d(s+1)$, то к шагу 8. Иначе сливаем два блока в один, присваиваем ему номер $s+1$ и для него определяем $d(s+1) = \frac{1}{2}\left(d(s)+d(s+1)\right)$.

8. Полагаем $s=s+1$. Если $s<m(l)$, то к шагу 7, иначе к шагу 9.

9. Если на шаге 7 происходило слияние каких-либо блоков, то $l=l+1$ и к шагу 6, иначе к шагу 10.

10. Пересчитаем координаты $\gamma_{ij}^{(k+1)} = d(s)$ при (i,j) в s-м блоке, $s=1,...,m(l)$, и по формуле Дж. Лингоса и Э. Роскама

$$\hat{x}_i^{(r,k+1)} = x_i^{(r,k)} - \frac{1}{n}\sum_{j=1}^{n}\left(1 - \frac{\gamma_{ij}^{(k+1)}}{d_{ij}^{(k)}}\right)\left(x_i^{(r,k)} - x_j^{(r,k)}\right),\ i=1,...,n,\quad r=1,...,p.$$

Если $d_{ij}^{(k)}=0$, то отношение $\frac{\gamma_{ij}^{(k+1)}}{d_{ij}^{(k)}}$ полагают равным 1.

Увеличиваем k: $k=k+1$. Переходим к шагу 2.

В результате работы алгоритма получаем числа $x_i^{(r,k)}$, $i=1,...,n$, $r=1,...,p$, которые суть наилучшие оценки координат изучаемых объектов в p-мерном пространстве с заданной точностью ε.

Шаги 3–9 называют неметрическим этапом алгоритма Торгерсона (см. в [4] другие варианты этого этапа). Шаг 10 алгоритма называется метрическим этапом. Формулы пересчета координат выписаны из условия уменьшения величины стресса.

Пример 8.1.

Исследовать данные по средней обеспеченности семей предметами быта, электроникой, средствами транспорта и дачами (всего 9 предметов) в 12 территориальных общностях (данные из массива msm.sta). В результате применения процедуры шкалирования территориальные общности должны рас-

положиться в двумерном геометрическом пространстве, построенном исходя из расстояний между ними по 9 переменным.

Решение.

Получим файл, в котором объектами будут территориальные общности, а переменными — обеспеченность домохозяйств предметами (v_1 — наличие телевизора, 0 — нет, 1 — есть; v_2 — наличие холодильника, 0 — нет, 1 — есть; v_3 — наличие стиральной машины, 0 — нет, 1 — есть; v_4 — наличие видеомагнитофона, 0 — нет, 1 — есть; v_5 — наличие микроволновой печи, 0 — нет, 1 — есть; v_6 — наличие компьютера, 0 — нет, 1 — есть; v_7 — наличие автомобиля, 0 — нет, 1 — есть; v_8 — наличие домашнего телефона, 0 — нет, 1 — есть; v_9 — наличие дачи, 0 — нет, 1 — есть;). Значения переменных — доли семей, обладающих этими предметами.

Затем данные агрегируем по территориальным общностям, сохранив доли семей, имеющих соответствующие предметы в новом файле. Затем вычисляем евклидовые расстояния между 12-ю территориями, используя 9 переменных. Получаем матрицу вида (8.1):

	1	2	3	4	5	6	7	8	9	10	11	12
1	0,00	0,46	0,46	0,38	0,25	0,27	0,37	0,34	0,48	0,16	0,22	0,39
2	0,46	0,00	0,82	0,18	0,68	0,63	0,65	0,68	0,84	0,52	0,42	0,70
3	0,46	0,82	0,00	0,71	0,29	0,25	0,32	0,20	0,21	0,37	0,49	0,23
4	0,38	0,18	0,71	0,00	0,57	0,53	0,57	0,58	0,72	0,43	0,38	0,61
5	0,25	0,68	0,29	0,57	0,00	0,18	0,36	0,22	0,29	0,20	0,36	0,31
6	0,27	0,63	0,25	0,53	0,18	0,00	0,33	0,15	0,32	0,20	0,35	0,27
7	0,37	0,65	0,32	0,57	0,36	0,33	0,00	0,25	0,35	0,29	0,27	0,17
8	0,34	0,68	0,20	0,58	0,22	0,15	0,25	0,00	0,25	0,21	0,35	0,15
9	0,48	0,84	0,21	0,72	0,29	0,32	0,35	0,25	0,00	0,38	0,49	0,24
10	0,16	0,52	0,37	0,43	0,20	0,20	0,29	0,21	0,38	0,00	0,23	0,28
11	0,22	0,42	0,49	0,38	0,36	0,35	0,27	0,35	0,49	0,23	0,00	0,35
12	0,39	0,70	0,23	0,61	0,31	0,27	0,17	0,15	0,24	0,28	0,35	0,00

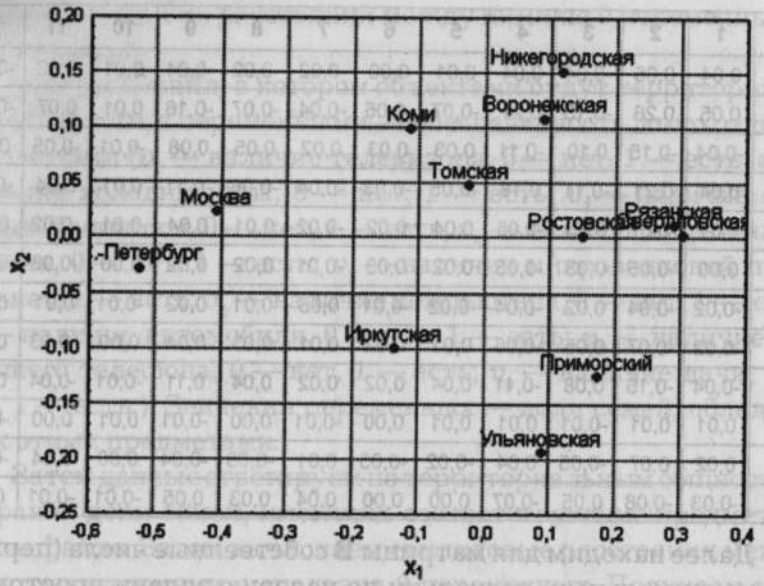
Применяя (8.4) для расчета элементов матрицы **B**, получаем

	1	2	3	4	5	6	7	8	9	10	11	12
1	0,04	0,05	-0,04	0,04	0,01	0,00	-0,02	-0,02	-0,04	0,01	0,02	-0,03
2	0,05	0,28	-0,15	0,21	-0,07	-0,05	-0,04	-0,07	-0,16	0,01	0,07	-0,08
3	-0,04	-0,15	0,10	-0,11	0,03	0,03	0,02	0,05	0,08	-0,01	-0,05	0,05
4	0,04	0,21	-0,11	0,18	-0,05	-0,03	-0,04	-0,06	-0,11	0,01	0,04	-0,07
5	0,01	-0,07	0,03	-0,05	0,04	0,02	-0,02	0,01	0,04	0,01	-0,02	0,00
6	0,00	-0,05	0,03	-0,03	0,02	0,03	-0,01	0,02	0,02	0,00	-0,03	0,00
7	-0,02	-0,04	0,02	-0,04	-0,02	-0,01	0,05	0,01	0,02	-0,01	0,01	0,04
8	-0,02	-0,07	0,05	-0,06	0,01	0,02	0,01	0,03	0,04	0,00	-0,03	0,03
9	-0,04	-0,16	0,08	-0,11	0,04	0,02	0,02	0,04	0,11	-0,01	-0,04	0,05
10	0,01	0,01	-0,01	0,01	0,01	0,00	-0,01	0,00	-0,01	0,01	0,00	-0,01
11	0,02	0,07	-0,05	0,04	-0,02	-0,03	0,01	-0,03	-0,04	0,00	0,04	-0,01
12	-0,03	-0,08	0,05	-0,07	0,00	0,00	0,04	0,03	0,05	-0,01	-0,01	0,05

Далее находим для матрицы **B** собственные числа (первые два максимальных, поскольку по условию задачи пространство координат двумерное) и соответствующие собственные векторы и по (8.5) получаем координатное представление 12 территорий:

Территория	Коми	С.-Петербург	Рязанская	Москва	Нижегородская	Воронежская
$X^{(1)}$	-0,11	-0,52	0,29	-0,41	0,12	0,09
$X^{(2)}$	0,10	-0,03	0,01	0,02	0,15	0,11
Территория	**Ульяновская**	**Ростовская**	**Свердловская**	**Томская**	**Иркутская**	**Приморский**
$X^{(1)}$	0,09	0,15	0,30	-0,03	-0,14	0,17
$X^{(2)}$	-0,19	0,00	0,00	0,05	-0,10	-0,12

Расположение точек на плоскости показано на рисунке:



Для интерпретации построенных шкал вычислим коэффициенты ранговой корреляции Спирмена (раздел 3.2) шкал и исходных переменных. Полученные значения сведены в таблицу:

	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
$x^{(1)}$	-0,27	-0,21	-0,51	-0,82	-0,63	-0,82	-0,62	-0,96	-0,61
	0,39	0,52	0,09	0,00	0,03	0,00	0,03	0,00	0,04
$x^{(2)}$	0,04	-0,23	-0,31	-0,43	-0,61	-0,29	-0,61	0,23	-0,38
	0,91	0,46	0,33	0,16	0,04	0,36	0,04	0,48	0,22

Под каждым коэффициентом ранговой корреляции указана его значимость. Для первой шкалы значимы на 1% уровне коэффициенты для переменных v_4, v_6, v_8, причем они имеют отрицательные знаки, для второй шкалы — на 5% уровне коэффициенты для переменных v_4 и v_6 и также отрицательны. Следовательно, первое измерение характеризует отсутствие средств

мультимедиа, второе — отсутствие сложной техники (автомобиля, микроволновки). Согласно рисунку $x^{(1)}$ имеет больший разброс, чем $x^{(2)}$ и, следовательно, объясняет большую долю расстояний объектов. По первой шкале в лидерах — города Москва и С.-Петербург — с учетом интерпретации шкалы эти территории являются наиболее развитыми по обеспеченности домохозяйств мультимедийной техникой, по второй — лидеры по обеспеченностью автотехникой — Приморский край, Ульяновская область.

В заключение примера отметим, что рассмотренный в главе 8 итерационный алгоритм МНШ дает результат с точки зрения оптимизации стресс-критерия и требует программных средств (в частности, пакета Statistica).

Вопросы и задачи

1. В чем суть задачи многомерного шкалирования?

2. Как решается задача метрического шкалирования по Торгерсону?

3. В чем отличие метрического шкалирования от неметрического?

4. Как строится матрица различий объектов?

5. Каков алгоритм решения задачи неметрического шаклирования?

6. Дана матрица различий по восьми странам. Вычислить матрицу скалярных произведений.

№ п.п.	Ангола	Аргентина	Австралия	Китай	Куба	Япония	США	Зимбабве
Ангола	0,00	1,41	1,00	1,00	1,41	1,41	1,73	0,71
Аргентина	1,41	0,00	1,00	1,73	1,41	1,41	1,00	1,41
Австралия	1,00	1,00	0,00	1,41	1,73	1,00	1,41	1,00
Китай	1,00	1,73	1,41	0,00	1,00	1,00	1,41	1,00

№ п.п.	Ангола	Аргентина	Австралия	Китай	Куба	Япония	США	Зимбабве
Куба	1,41	1,41	1,73	1,00	0,00	1,41	1,00	1,41
Япония	1,41	1,41	1,00	1,00	1,41	0,00	1,00	1,41
США	1,73	1,00	1,41	1,41	1,00	1,00	0,00	1,73
Зимбабве	0,71	1,41	1,00	1,00	1,41	1,41	1,73	0,00

7. Примените метод главных компонент для извлечения трех компонент из вычисленной в задаче 6 матрицы скалярных произведений. Вычислите координаты стран в трехмерном пространстве.

Литература

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики: Учебник для вузов. — М.: ЮНИТИ, 1998. — 1022 с.

2. Айвазян С.А., Бухштабер В.М., Енюков И.С. и др. Прикладная статистика. Классификация и снижение размерности. Справ. изд. — М.: Финансы и статистика, 1989. — 607 с.

3. Айвазян С.А., Мхитарян В.С. Прикладная статистика в задачах и упражнениях: Учебник для вузов. — М.: ЮНИТИ-ДАНА, 2001. — 270 с.

4. Дэйвисон М. Многомерное шкалирование: методы наглядного представления данных/Пер. с англ. — М.: Финансы и статистика, 1988. — 254 с.

5. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: Учебник. — М.: Финансы и статистика, 2000. — 352 с.

6. Дубров А.М. Компонентный анализ и эффективность в экономике: Учеб. пособие. — М.: Финансы и статистика, 2002. — 352 с.

7. Кремер Н.Ш. Теория вероятностей и математическая статистика: Учебник для вузов. — М.: ЮНИТИ-ДАНА, 2004. — 573 с.

8. Сажин Ю.В., Басова В.А., Катынь А.В. Многомерные статистические методы анализа экономических процессов: Учеб. пособие. — Саранск: Изд-во Мордовского ун-та, 2000. — 88 с.

9. Сошникова Л.А., Тамашевич В.Н., Уебе Г., Шефер М. Многомерный статистический анализ в экономике: Учеб. пособие для вузов/ Под ред. проф. В.Н.Тамашевича. — М.: ЮНИТИ-ДАНА, 1999. — 598 с.

10. Средние классы в России: экономические и социальные стратегии/Под ред. Т. Малевой. Моск. Центр Карнеги. — М.: Гендальф, 2003. — 506 с.

11. Факторный, дискриминантный и кластерный анализ: Пер. с англ./Дж.-О. Ким, Ч.У. Мьюллер и др.;// Под ред. И.С. Енюкова. — М.: Финансы и статистика, 1989. — 215 с.

12. *Харман Г.* Современный факторный анализ/Пер. с англ. — М.: Статистика, 1972. — 486 с.

13. Экономико-математический энциклопедический словарь/ Гл. ред. Данилов-Данильян. — М.: Большая Российская Энциклопедия: Изд. дом «ИНФРА-М»,2003. — 688 с.

14. Johnson R.A., Wichern D.W. Applied multivariate statistical analysis. Prentice Hall. New Jersey, 2002.

15. Hair J.F. Multivariate data analysis. Prentice Hall. New Jersey, 1998.

Ресурсы Интернет

1. www.socpol.ru Независимый институт социальной политики. Сайт содержит великолепный архив социально-экономических данных, которые можно использовать как в учебных, так и в научных целях.

2. www.eerc.ru Консорциум экономических исследований и образования. На сайте доступно большое количество публикаций по исследованиям в области экономики.

3. www.cemi.rssi.ru Сайт Центрального экономико-математического института РАН, также см. www.nes.ru — Российская экономическая школа.

4. www.exponenta.ru Сайт, посвященный применению пакетов прикладных программ в учебных целях. Имеется отдельная страничка по пакету Statistica.

5. www.statsoft.ru, www.statsoft.ru/home/portal/ и www.statsoft.com Сайт разработчика Statistica. Русскоязычная версия содержит также информацию по практике использования методов многомерного анализа в экономике. Имеется большое количество справочной информации по Statistica и многомерным методам.

6. www.biometrica.tomsk.ru/list/statleo.htm Ссылки на статистические ресурсы.

7. repec.org База данных различных материалов (статьи, препринты и программы) по экономике.

8. teorver-online.narod.ru/oknige.html Учебник по теории вероятностей и математической статистике.

9. www.ecsocman.edu.ru Образовательный портал по экономике.

10. www.cir.ru Университетская информационная система — база электронных ресурсов для исследований и образования в области экономики и других гуманитарных наук.

Краткий терминологический словарь

Варимакс (varimax) — метод получения ортогонального решения, который сводится к упрощению факторной структуры с использованием критерия минимизации дисперсии столбца матрицы факторного отображения.

Вторичные оси (reference axes) — оси, ортогональные первичным факторам; вводятся для упрощения косоугольного вращения.

Выделение факторов (extraction of factors) — первоначальный этап факторного анализа; ковариационная матрица воспроизводится посредством небольшого числа скрытых факторов или компонент.

Генеральная совокупность (population) — множество всех мыслимых наблюдений, которые могли бы быть произведены при данном комплексе условий.

Главные компоненты (principal components) — линейная комбинация наблюдаемых переменных, обладающая свойством ортогональности; первая главная компонента воспроизводит наибольшую долю дисперсии экспериментальных данных; вторая — следующую по величине долю и т. д.; главные компоненты часто считаются общими факторами, но более корректно предположение, что они противоположны им, поскольку общие факторы являются гипотетическими.

Главных осей метод (principal axis factoring) — метод получения первоначального факторного решения, при использовании которого редуцированная корреляционная матрица подвергается последовательной декомпозиции; метод главных осей с итерациями по общности эквивалентен методу наименьших квадратов.

Дискриминантныйанализ(discriminantanalyses) — раздел статистического анализа, посвященный получению правил классификации, наблюдений (объектов) в один из нескольких описанных некоторым образом классов.

Дискриминантная функция (discriminant function) — статистика, служащая для построения правила классификации объектов по заданным группам.

Дискриминантная функция Фишера (Fisher discriminant function) — линейная комбинация исходных (дискриминирующих) переменных, которая выбирается так, чтобы расхождение между классами объектов было максимально возможным.

Значение фактора (factor score) — оценка скрытого фактора в терминах наблюдаемых переменных; в факторном анализе имеет второстепенное значение.

Иерархический агломеративный (дивизимный) метод (agglomerative (divisive) hierarchical method) — метод кластерного анализа, суть которого в последовательном объединении (разделении) групп объектов сначала самых близких (далеких), а затем все более отдаленных (близких) друг от друга. Позволяет получить стратификационную структуру классифицируемой совокупности в форме дендрограммы.

Категоризованная (дискретная) переменная (categorical variable) — переменная, измеренная в номинальной или ранговой шкале. Значения такой переменной часто называют градациями. Множество объектов (статистических единиц), соответствующих одной и той же градации, называют категорией объектов.

Квартимакс (quartimax) — критерий получения ортогонального решения; сводится к упрощению описания строк матрицы факторного отображения.

Классификация (classification) — отнесение объектов, элементов некоторого множества к тому или иному классу (подмножеству, элементы которого характеризуются неким существенным признаком или группой существенных признаков) или разделение совокупности объектов на однородные в определенном смысле группы.

Кластер (cluster) — это группа элементов, обладающих каким-то общим свойством и находящихся на небольшом расстоянии друг от друга.

Кластерный анализ (cluster analysis) — математически-формализованная процедура разбиения анализируемой совокупности объектов на некоторое число (заранее известное или нет) однородных в определенном смысле классов в условиях отсутствия обучающих выборок.

Ковариацийанализ(covariance-structureanalysis) — метод анализа, в котором: 1) наблюдаемые коэффициенты ковариаций описываются в рамках общей модели, включающей гипотетические факторы и наблюдаемые переменные; 2) исследователь затем определяет соответствующие значения, оценивая адекватность этого определения по отношению к структуре выборочных ковариаций.

Конфирматорный факторный анализ (confirmatory factor analysis) — факторный анализ, в котором проверяются гипотезы о числе факторов и их нагрузках.

Корреляционный анализ (analysis of correlation) — совокупность основанных на теории корреляции методов обнаружения корреляционной зависимости между случайными величинами или признаками.

Корреляция (correlation) — мера зависимости между двумя переменными.

Косоугольное вращение (oblique rotation) — преобразование, с помощью которого получается простая структура; факторы вращаются без наложения условия ортогональности, и результирующие факторы, вообще говоря, коррелируют друг с другом.

Косоугольные факторы (oblique factors) — факторы, которые коррелируют друг с другом; получаются в результате косоугольного вращения.

Линейная комбинация (linear combination) — сумма, в которую переменные входят с постоянными весами.

Максимального правдоподобия метод (maximum likelihood) — метод статистического оценивания, в котором определяется значение переменных генеральной совокупности с использованием выборочного распределения.

Метод k-средних (k-means method) — предназначен для разбиения многомерных наблюдений на заданное число, однородных в смысле геометрической взаимной близости элементов, принадлежащих к одному классу, минимизируя усредненную меру внутриклассового разброса наблюдений.

Многомерная генеральная совокупность (multivariate population) –совокупность, когда на каждом из ее объектов регистрируются значения набора признаков $x_1,...,x_p$.

Многомерное наблюдение (multivariate observation) — результат регистрации значений многомерного признака на каком-либо статистически обследованном объекте.

Многомерное шкалирование (multidimensional scaling) — математические методы, позволяющие по заданной информации о мерах различия (близости) между объектами рассматриваемой совокупности приписывать каждому из этих объектов вектор характеризующих его количественных показателей.

Многомерный признак (multivariate attribute) — вектор показателей, среди которых могут быть количественные, порядковые и номинальные.

Многомерный статистический анализ (multivariate statistical analysis) — раздел математической статистики, объединяющий методы изучения статистических данных, которые являются значениями многомерных качественных или количественных признаков.

Многомерная функция распределения (multivariate distribution function) — функция, задающая совместное распределе-

ние вероятностей нескольких случайных величин X, Y,...; для любого набора значений x, y, ... она равна вероятности того, что случайная величина X меньше или равна x и при этом случайная величина Y меньше или равна y, и т.д.

Множественный коэффициент корреляции (multiple correlation coefficient) — измеряет степень тесноты статистической связи между некоторым (результирующим) показателем, с одной стороны, и совокупностью других (объясняющих) переменных — с другой.

Модель смеси распределений (mixture of distributions) — описывает закон распределения вероятностей в смеси из нескольких одномодальных генеральных совокупностей, имеющих в общей генеральной совокупности удельные веса и функции плотности.

Монте-Карло метод (Monte Carlo experiment) — методика статистического моделирования выборочных характеристик.

Наименьших квадратов метод (least-squares solution) — решение, для которого минимизируется сумма квадратов отклонений между наблюдаемыми и предполагаемыми значениями.

Непараметрический дискриминантный анализ (non-parametric discriminant analyses) — статистическая реализация байесовского правила классификации в ситуации, когда законы распределения многомерного признака внутри классов неизвестны.

Облимакс (oblimax) — критерий получения косоугольного решения; эквивалентен критерию квартимакс при ортогональном вращении.

Общий фактор (common factor) — неизмеряемая (гипотетическая) скрытая величина, которая учитывает корреляцию по крайней мере между двумя наблюдаемыми переменными.

Общность (communality) — доля дисперсии наблюдаемых переменных, обусловленная общими факторами; в модели с ортогональными факторами она равна сумме квадратов факторных нагрузок.

Оптимальное (байесовское) правило классификации (Bayes decision rule) — процедура классификации, основанная на идее наибольшего правдоподобия — объект следует отнести к тому классу, в рамках которого он выглядит наиболее правдоподобным, т.е. такая процедура минимизирует вероятность принятия ошибочного решения.

Ортогональное вращение (orthogonal rotation) — преобразование, с помощью которого получается простая структура при выполнении ограничения ортогональности (некоррелированности) факторов; факторы, выделяемые с помощью этого вращения, по определению не коррелированы.

Ортогональные факторы (orthogonal factors) — факторы, которые не коррелируют друг с другом; получаются при ортогональном вращении.

Основание (критерий) классификации (classification criterion) — существенный признак, по которому проводится классификация.

Параметрический дискриминантный анализ (parametric discriminant analyses) — статистическая реализация байесовского правила классификации в ситуации, когда законы распределения многомерного признака внутри классов заданы с точностью до неизвестного параметра (параметров).

Признак (attribute) — существенная для данного исследования характеристика (свойство) объекта.

Разведочный факторный анализ (exploratory factor analysis) — факторный анализ, который используется при исследовании скрытой факторной структуры без предположения о числе факторов и их нагрузок.

Ранговая корреляция (rang correlation) — мера зависимости между признаками, выраженным в порядковой шкале.

Редуцированная корреляционная матрица (adjusted correlation matrix) — корреляционная матрица, в которой элемен-

ты главной диагонали соответствуют общностям: корреляционные или ковариационные матрицы, которыми пользуются перед выделением факторов.

Специфичность (specific component) — доля дисперсии наблюдаемой переменной, соответствующая специфичному фактору; применяется для обозначения части характерности, получаемой при исключении дисперсии ошибки.

Сумма квадратов отклонений (variation) — мера разброса переменной; сумма квадратов отклонений от среднего.

Таблица сопряженности (contingency table) — задает распределение объектов по категориям двух или нескольких нечисловых признаков.

Условное распределение вероятностей (conditional probability distribution) — многомерное распределение случайных величин, которое получается, когда значения одной или нескольких из них фиксированы

Факторы (factors) — гипотетические, непосредственно неизмеряемые, скрытые переменные, в терминах которых описываются измеряемые переменные; часто подразделяются на характерные и общие.

Факторная нагрузка (factor loading) — общий термин, означающий коэффициенты матрицы факторного отображения или структуры.

Факторного отображения матрица (factor pattern matrix) — матрица коэффициентов, в которой столбцы соответствуют общим факторам, а строки — наблюдаемым переменным; элементы матрицы факторного отображения представляют собой коэффициенты регрессии для общих факторов при условии, что наблюдаемые переменные являются линейной комбинацией факторов; для ортогонального решения матрица отображения содержит коэффициенты корреляции между переменными и факторами.

Факторная сложность переменной (factorial complexity) — характеристика наблюдаемой переменной представляет собой число общих факторов с ненулевыми нагрузками для данной переменной.

Факторной структуры матрица (factor structure matrix) — матрица коэффициентов корреляции между переменными и факторами; в случае некоррелированных (ортогональных) факторов совпадает с матрицей факторного отображения.

Факторной причинности принцип (postulate of factorial causation) — предположение о том, что наблюдаемые переменные являются линейной комбинацией скрытых факторов и что ковариации между наблюдаемыми переменными воспроизводятся с помощью общих факторов.

Характерность (unique component) — доля дисперсии наблюдаемой переменной, не связанная с общими факторами и свойственная именно данной переменной; она часто разделяется на специфичность и дисперсию ошибки.

Характерный фактор (unique factor) — фактор, влияющий только на данную переменную; часто относится ко всем независимым факторам (включая ошибку измерений), характерным только для данной переменной.

Центроид (centroid) — точка (центр), координатами которой являются средние значения по каждой из размерностей.

Шкала (scale) — система чисел или иных элементов, принятых для оценки или измерения каких-либо величин.

Экономии принцип (postulate of parsimony) — состоит в том, что из двух конкурирующих моделей выбирается наиболее простая.

Источник данных

В качестве одного из источников данных воспользуемся информационной базой из [10], которая содержит данные специально организованного выборочного обследования. Выборка строилась как репрезентирующая все население России и отдельные типы поселений в разрезе:

- областных центров (включая Москву и С.-Петербург);
- городов областного подчинения;
- сельских поселений.

Системные смещения выборки определялись следующими параметрами:

- отказом от наблюдений за представителями социального «дна» (нищие, бомжи, беспризорники и прочие асоциальные группы);
- отказом от наблюдений за представителями высших социальных слоев (элит и их домохозяйств).

Причина такой заданной априори смещенности состоит в невозможности обеспечить присутствие в поселенческой случайной выборке представителей этих социальных слоев. Они традиционно изучаются по специальным методикам.

В качестве единицы наблюдения в исследовании принято домохозяйство. Под домохозяйством понимается группа лиц, не обязательно связанных родственными и свойскими отношениями, но обязательно проживающих совместно и имеющих общий бюджет. Этим утверждается, что домохозяйство является системной единицей социального поведения. В домохозяйстве аккумулируются социальные ресурсы его членов (экономические, интеллектуальные, имиджевые, коммуникационные и пр.). Они расходуются в интересах домохозяйства в целом и отдельных его представителей.

В исследовании респондентом являлось лицо, случайно отобранное из членов домохозяйства в возрасте 18 лет и старше, представляющее как специальным образом отобранное домохозяйство, так и самого себя.

Поэтому инструментарий исследования состоял из индикаторов двух типов:

- относящихся к домохозяйству или всем его членам;
- касающихся только случайно отобранного респондента.

В тех случаях, когда случайно отобранный респондент не мог ответить на вопросы, относящиеся к домохозяйству в целом, к обследованию привлекался наиболее информированный член домохозяйства. В результате в качестве исходных параметров получены:

- широкий круг домохозяйственных характеристик, репрезентирующих на общероссийском и поселенческом уровнях российские домохозяйства в целом;
- ограниченный, но достаточно информативный набор данных о каждом члене домохозяйства, представляющий население в целом;
- обширная база данных по случайно отобранным взрослым членам домохозяйств, репрезентирующая население России в возрасте 18 лет и старше.

Подробное описание дизайна выборки и анкеты обследования имеется в [10]. Файл данных msm.sta содержит 5028 индивидуальных наблюдений и следующие переменные, используемые нами для примеров и упражнений:

Переменная	Среднее	Стандартное отклонение
Пол (0 — мужской, 1 — женский)	0,61	0,49
Возраст	47,54	17,26
Состояние в браке (1 — женат/замужем, 2 — разведен/разведена, 3 — вдова/вдовец, 4 — никогда не состояли в браке)	1,79	1,11
Число лет образования	12,16	3,84
Владеете ли вы английским языком (1 — не владею, 2 — читаю, перевожу со словарем, 3 — читаю, могу объясняться, 4 — владею свободно)	1,27	0,61

Переменная	Среднее	Стандартное отклонение
Отрасль занятости (1 — топливно-энергетический комплекс, 2 — ВПК, 3 — др. отрасли промышленности, 4 — строительство, транспорт, 5 — связь, информационные технологии, 6 — сельское и лесное хозяйство, 7 — оптовая торговля, 8 — розничная торговля, общественное питание, бытовое обслуживание, 9 — жкх, 10 — финансы, кредитование, страхование, 11 — наука, 12 — образование, 13 — здравоохранение, 14 — культура, искусство, журналистика, 15 — госуправление, 16 — юридические услуги, 17 — прочие услуги, 18 — армия, милиция, ФСБ, охрана, 19 — другое, 20 — затрудняюсь ответить)	8,23	4,94
Сколько процентов занятости занимает основная работа?	84,47	32,44
Общая площадь жилья	47,63	17,87
Среднедушевой доход	1585,90	2485,42
Наличие сбережений (0 — нет, 1 — да)	0,25	0,43
Наличие доходов (0 — нет, 1 — да)	0,32	0,47
Имущественная обеспеченность (0 — нет, 1 — да)	0,37	0,48
Наличие жилья (0 — нет, 1 — да)	0,34	0,48
Наличие скота (0 — нет, 1 — да)	0,08	0,28
Наличие земли (0 — нет, 1 — да)	0,04	0,19
Тип поселения (1 — областной центр, 2 — районный центр, 3 — село)	1,65	0,77

Некоторые сведения из линейной алгебры*

1. Векторное пространство.

Определение. Вещественным векторным пространством называется множество L, элементы которого называются векторами, удовлетворяющее следующим условиям (аксиомам).

1. Определена операция сложения векторов, результатом которой является вектор: $\mathbf{a}, \mathbf{b} \in L \Rightarrow \mathbf{a}+\mathbf{b} \in L$.

2. $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$ для всех $\mathbf{a}, \mathbf{b} \in L$ (коммутативность).

3. $\mathbf{a} + (\mathbf{b} + \mathbf{c}) = (\mathbf{a} + \mathbf{b}) + \mathbf{c}$ для всех $\mathbf{a}, \mathbf{b}, \mathbf{c} \in L$ (ассоциативность).

4. Существует нулевой вектор $\mathbf{0}$, такой, что $\mathbf{0} + \mathbf{a} = \mathbf{a} + \mathbf{0} = \mathbf{a}$ для любого $\mathbf{a} \in L$.

5. Для всякого вектора $\mathbf{a} \in L$ и вещественного числа $\alpha \in R$ определено их произведение $\alpha\mathbf{a} \in L$.

6. $(\alpha+\beta)\mathbf{a} = \alpha\mathbf{a}+\beta\mathbf{a}$ для всех $\alpha, \beta \in R$ и $\mathbf{a} \in L$.

7. $\alpha(\mathbf{a}+\mathbf{b}) = \alpha\mathbf{a}+\alpha\mathbf{b}$ для всех $\alpha \in R$ и $\mathbf{a}, \mathbf{b} \in L$.

8. $0\mathbf{a} = \mathbf{0}$ для всех $\mathbf{a} \in L$.

9. $1\mathbf{a} = \mathbf{a}$ для всех $\mathbf{a} \in L$.

2. Векторное пространство R^n.

Элементами (точками, векторами) вещественного векторного пространства R^n являются векторы-столбцы, состоящие из n вещественных чисел $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$; операции сложения и умножения на число определены следующим образом:

$$\mathbf{a}+\mathbf{b} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} a_1+b_1 \\ \vdots \\ a_n+b_n \end{pmatrix}; \quad \alpha\mathbf{a} = \begin{pmatrix} \alpha a_1 \\ \vdots \\ \alpha a_n \end{pmatrix}.$$

* Воспроизводится по учебнику: *Магнус Я.Р., Катышев П.К., Пересецкий А.А.* Эконометрика. Начальный курс: Учеб. 6-е изд. — М.: Дело, 2004. — 576 с.

Нулевой вектор имеет все координаты, равные 0.

3. Линейная зависимость.

Определение. Векторы $\mathbf{a}_1,...,\mathbf{a}_k$ называются *линейно независимыми*, если из того, что $\alpha_1\mathbf{a}_1+\alpha_2\mathbf{a}_2+...\alpha_k\mathbf{a}_k=\mathbf{0}$, $\alpha_i\in R$, следует, что все $\alpha_i=0$.

Векторы $\mathbf{a}_1,...,\mathbf{a}_k$ называются линейно зависимыми, если существует набор α_i, где хотя бы одно α_i отлично от нуля, удовлетворяющий вышеуказанному условию.

4. Линейное подпространство.

Определение. Линейным подпространством линейного пространства L называется подмножество K векторов пространства L, замкнутое относительно операций сложения и умножения на число, т. е. из того, что векторы $\mathbf{a}, \mathbf{b}\in K$, следует, что $\mathbf{a}, \mathbf{b}\in K$ и $\alpha\mathbf{a}\in K$.

Определение. Множество всех *линейных комбинаций* векторов $\mathbf{a}_1,...,\mathbf{a}_k\in L$ $\alpha_1\mathbf{a}_1+\alpha_2\mathbf{a}_2+...+\alpha_k\mathbf{a}_k, \alpha_i\in R$ называется *пространством, порожденным векторами* $\mathbf{a}_1,...,\mathbf{a}_k$.

Если линейное подпространство K векторного пространства L не совпадает с ним, то его часто называют *гиперплоскостью*.

5. Базис. Размерность.

Определение. Набор векторов $\mathbf{a}_1,...,\mathbf{a}_n\in L$ называется базисом пространства L, если выполняются условия: векторы $\mathbf{a}_1,...,\mathbf{a}_n$ линейно независимы, пространство, порожденное векторами $\mathbf{a}_1,...,\mathbf{a}_n$, совпадает с L.

Предложение. Все базисы векторного пространства L содержат одно и то же число векторов, которое называется размерностью $\dim(L)$ векторного пространства L.

Пример. Размерность R^n равна $\dim(R^n)=n$.

Предложение. Любой вектор $\boldsymbol{a}$ линейного пространства можно *единственным* способом *разложить по базису*, т. е. представить в виде линейной комбинации базисных векторов: $\mathbf{a}=\alpha_1\mathbf{e}_1+\alpha_2\mathbf{e}_2+...+\alpha_n\mathbf{e}_n, \alpha_i\in R$.

6. Матрицы.

Определение. $m\times n$ матрицей называется прямоугольная таблица чисел, где первый индекс означает номер строки, а второй — номер столбца.

$$\mathbf{A}=(a_{ij})=\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}.$$

Матрица $n\times 1$ называется *вектором-столбцом*; матрица $1\times n$ — *вектором-строкой*; 1×1 матрица называется *скалярной* матрицей. Далее, если не оговорено противное, мы будем везде рассматривать вектор как вектор-столбец.

Квадратная матрица, в которой все элементы, не лежащие на *главной диагонали*, равны 0, называется *диагональной*:

$$\Lambda=\begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \lambda_n \end{pmatrix}.$$

Диагональная матрица $\mathbf{I}_n$, у которой все диагональные элементы равны 1, называется *единичной* (индекс n здесь обозначает размерность матрицы и может быть опущен).

Нулевой матрицей называется матрица, состоящая из одних нулей.

7. Операции с матрицами.

Определение. Две матрицы $\mathbf{A}$ и $\mathbf{B}$ *равны*, если совпадают их размерности и равны их соответствующие элементы.

Определение. Суммой двух матриц $\mathbf{A}=(a_{ij})$ и $\mathbf{B}=(b_{ij})$ размерностей $m\times n$ называется матрица $\mathbf{A}+\mathbf{B}=\mathbf{C}=(c_{ij})$ размерности $m\times n$ с элементами $c_{ij}=a_{ij}+b_{ij}$, т. е. при сложении матриц складываются соответствующие элементы.

Определение. Произведением $m\times n$ матрицы $\mathbf{A}=(a_{ij})$ *на число* $\alpha\in R$ называется матрица $\alpha\mathbf{A}=\mathbf{C}=(c_{ij})$ размерности $m\times n$ с элементами $c_{ij}=\alpha a_{ij}$, т. е. при умножении матрицы на число все элементы матрицы умножаются на это число.

Предложение. Операция сложения матриц удовлетворяет следующим свойствам:

$\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$,

$(\mathbf{A}+\mathbf{B})+\mathbf{C}=\mathbf{A}+(\mathbf{B}+\mathbf{C})$,

$\alpha(\mathbf{A}+\mathbf{B})=\alpha\mathbf{A}+\alpha\mathbf{B}$,

$\mathbf{A}+\mathbf{0}=\mathbf{A}$.

Определение. *Транспонированной* матрицей называется матрица, у которой строки и столбцы поменялись местами, а именно: для $m\times n$ матрицы $\mathbf{A}=(a_{ij})$ транспонированной является $n\times m$ матрица $\mathbf{A}^T=(a_{ji})$. Например, $\begin{pmatrix}1&4&-3\\2&5&0\end{pmatrix}^T=\begin{pmatrix}1&2\\4&5\\-3&0\end{pmatrix}$.

Предложение. Свойства операции транспонирования матриц:

$(\mathbf{A}+\mathbf{B})^T=\mathbf{A}^T+\mathbf{B}^T$, $(\mathbf{A}^T)^T=\mathbf{A}$.

Определение. Пусть мы имеем матрицы $\mathbf{A}$ размерности $m\times n$ и $\mathbf{B}$ размерности $n\times k$, т. е. число столбцов у матрицы $\mathbf{A}$ равно числу строк у матрицы $\mathbf{B}$. *Произведением двух матриц* $\mathbf{A}$, $\mathbf{B}$ называется $m\times k$ матрица $\mathbf{C}=\mathbf{AB}$, элементы которой определяются следующим образом: $c_{ij}=\sum_{s=1}^{n}a_{is}b_{sj}$, $i=1,\dots,m, j=1,\dots,k$.

Пример.

$$\mathbf{AB}=\begin{pmatrix}1&0\\0&1\\2&3\end{pmatrix}\begin{pmatrix}1&-1&2\\0&3&4\end{pmatrix}=\begin{pmatrix}1\cdot1+0\cdot0&1\cdot(-1)+0\cdot3&1\cdot2+0\cdot4\\0\cdot1+1\cdot0&0\cdot(-1)+1\cdot3&0\cdot2+1\cdot4\\2\cdot1+3\cdot0&2\cdot(-1)+3\cdot3&2\cdot2+3\cdot4\end{pmatrix}=$$

$$=\begin{pmatrix}1&-1&2\\0&3&4\\2&7&16\end{pmatrix}.$$

Операция произведения матриц вообще говоря *некоммутативна*: $\mathbf{AB}\neq\mathbf{BA}$. Более того, $\mathbf{AB}$ может быть определено, а $\mathbf{BA}$ — не определено вовсе.

Определение. *Скалярным произведением* двух векторов $\mathbf{a}$, $\mathbf{b}$ размерности n называется число, равное $\mathbf{a}^T\mathbf{b}=\mathbf{b}^T\mathbf{a}=a_1b_1+a_2b_2+\dots+a_nb_n$.

Пример.

$$\mathbf{a}^T\mathbf{b}=\begin{pmatrix}1\\2\\3\end{pmatrix}^T\begin{pmatrix}2\\0\\-1\end{pmatrix}=1\cdot2+2\cdot0+3\cdot(-1)=-1.$$

Замечание 1. Элемент с координатами i, j в произведении матриц $\mathbf{AB}$ равен скалярному произведению i-го вектора-строки матрицы $\mathbf{A}$ на j-й вектор-столбец матрицы $\mathbf{B}$.

Замечание 2. Важным частным случаем произведения матриц является произведение квадратной $n\times n$ матрицы $\mathbf{A}$ на вектор $\mathbf{b}$. Например,

$$\mathbf{Ab}=\begin{pmatrix}1&2&3\\1&0&1\\2&4&6\end{pmatrix}\begin{pmatrix}b_1\\b_2\\b_3\end{pmatrix}=\begin{pmatrix}1\cdot b_1+2\cdot b_2+3\cdot b_3\\1\cdot b_1+0\cdot b_2+1\cdot b_3\\2\cdot b_1+4\cdot b_2+6\cdot b_3\end{pmatrix}=$$

$$=b_1\begin{pmatrix}1\\1\\2\end{pmatrix}+b_2\begin{pmatrix}2\\0\\3\end{pmatrix}+b_3\begin{pmatrix}3\\1\\6\end{pmatrix}.$$

Как видно из примера, вектор $\mathbf{Ab}$ является *линейной комбинацией* столбцов матрицы $\mathbf{A}$ с коэффициентами b_i.

Аналогично при умножении матрицы $\mathbf{A}$ на вектор-строку (слева) $\mathbf{b}^T\mathbf{A}$ мы получаем вектор-строку, являющуюся линейной комбинацией *строк* матрицы $\mathbf{A}$ с коэффициентами b_i.

Предложение. Свойства операции умножения матриц:

$\mathbf{AI}=\mathbf{A}$, $\mathbf{IA}=\mathbf{A}$, ($\mathbf{I}$ — единичная матрица подходящей размерности),

$\mathbf{A}(\mathbf{B}+\mathbf{C})=\mathbf{AB}+\mathbf{AC}$,

$(\mathbf{A}+\mathbf{B})\mathbf{C}=\mathbf{AC}+\mathbf{BC}$,

$\mathbf{A}(\mathbf{BC})=(\mathbf{AB})\mathbf{C}$,

$(\mathbf{AB})^T=\mathbf{B}^T\mathbf{A}^T$, $(\mathbf{ABC})^T=\mathbf{C}^T\mathbf{B}^T\mathbf{A}^T$

$\mathbf{A0}=\mathbf{0}$.

8. Инварианты матриц: след, определитель.

В дальнейшем часто используются две числовые функции, определенные только для *квадратных* матриц: *след* матрицы и *определитель* (*детерминант*) матрицы.

Определение. След (*trace*) матрицы равен сумме ее диагональных элементов:

$\mathrm{tr}(\mathbf{A}) = a_{11} + a_{22} + \ldots + a_{nn} = \sum a_{ii}$.

Предложение. Свойства следа матриц:

$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$,

$\mathrm{tr}(\boldsymbol{I}_n) = n$,

$\mathrm{tr}(\alpha\mathbf{A}) = \alpha\,\mathrm{tr}(\mathbf{A})$,

$\mathrm{tr}(\mathbf{A}^{\mathrm{T}}) = \mathrm{tr}(\mathbf{A})$,

$\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$.

Если $\mathbf{a}$ — вектор-столбец, то $\mathrm{tr}(\mathbf{a}\mathbf{a}^{\mathrm{T}}) = \mathrm{tr}(\mathbf{a}^{\mathrm{T}}\mathbf{a}) = \mathbf{a}^{\mathrm{T}}\mathbf{a}$ ($\mathbf{a}^{\mathrm{T}}\mathbf{a}$ — скалярный квадрат вектора $\mathbf{a}$).

Определение. Определителем (*детерминантом*) $\det(\mathbf{A}) = |\mathbf{A}|$ квадратной $n \times n$ матрицы $\mathbf{A}$ называется числовая функция матриц, удовлетворяющая следующим условиям:

1. $n = 1$, $\det(\mathbf{A}) = a_{11}$;

2. «*разложение определителя по строке*» при $n > 1$:

$|\mathbf{A}| = \sum_{j=1}^{n} a_{ij}(-1)^{i+j}|\mathbf{A}_{ij}|$, где $\mathbf{A}_{ij}$ — $(n-1)\times(n-1)$ матрица, получающаяся из исходной вычеркиванием i-й строки и j-го столбца. Определитель $|A_{ij}|$ называется *минором* порядка $n-1$ матрицы $\mathbf{A}$.

Условия 1, 2 дают рекуррентное определение детерминанта матрицы. Для малых размерностей удобно пользоваться формулами:

$$n = 2:\ \det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21};$$

$$n = 3:\ \det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} -$$

$$-a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32}.$$

Предложение. Свойства определителя матриц:

$\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$,

$\det(\mathbf{I}_n) = 1$,

$\det(\alpha\mathbf{A}) = \alpha^n\det(\mathbf{A})$,

$\det(\mathbf{A}^{\mathrm{T}}) = \det(\mathbf{A})$,

при перестановке двух строк (столбцов) определитель меняет знак,

определитель равен 0, если в нем есть две одинаковые строки (столбца),

определитель не меняется, если к одной строке (столбцу) матрицы добавить линейную комбинацию других строк (столбцов),

определитель равен 0 тогда и только тогда, когда строки (столбцы) линейно зависимы.

9. Ранг матрицы.

Пусть $\mathbf{A}$ — $m \times n$ матрица (не обязательно квадратная).

Определение. Рангом по строкам матрицы $\mathbf{A}$ называется размерность линейного подпространства в R^n, порожденного m векторами-строками матрицы $\mathbf{A}$.

Определение. Рангом по столбцам матрицы $\mathbf{A}$ называется размерность линейного подпространства в R^m, порожденного n векторами-столбцами матрицы $\mathbf{A}$.

Определение. Рангом по минорам матрицы $\mathbf{A}$ называется наибольший порядок ненулевого минора матрицы $\mathbf{A}$. (Минор порядка k матрицы — определитель квадратной $k \times k$ матрицы, получающейся из исходной матрицы вычеркиванием некоторого количества строк и столбцов.)

Предложение. Все три приведенных выше определения дают одно и то же число, называемое *рангом матрицы:* $\mathrm{rang}(\mathbf{A})$.

Предложение. Свойства ранга матрицы:

$\mathrm{rang}(\mathbf{A}) \le \min(m, n)$,

$\mathrm{rang}(\mathbf{AB}) \le \min(\mathrm{rang}(\mathbf{A}), \mathrm{rang}(\mathbf{B}))$,

если $\mathbf{B}$ — $n \times n$ квадратная матрица ранга n, то $\mathrm{rang}(\mathbf{AB}) = \mathrm{rang}(\mathbf{A})$,

если $\mathbf{B}$ — $m \times m$ квадратная матрица ранга m, то $\mathrm{rang}(\mathbf{BA}) = \mathrm{rang}(\mathbf{A})$,

$\mathrm{rang}(\mathbf{A}) = \mathrm{rang}(\mathbf{AA}^{\mathrm{T}}) = \mathrm{rang}(\mathbf{A}^{\mathrm{T}}\mathbf{A})$, причем $\mathbf{AA}^{\mathrm{T}}$ — $m \times m$ матрица, а $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ — $n \times n$ матрица.

10. Обратная матрица.

Пусть $\mathbf{A}$ — квадратная $n\times n$ матрица.

Определение. Матрица $\mathbf{A}$ называется *невырожденной*, если она имеет максимальный возможный ранг: $\mathrm{rang}(\mathbf{A}) = n$.

Определение. Матрицей, *обратной* к матрице $\mathbf{A}$, называется матрица, обозначаемая $\mathbf{A}^{-1}$, такая, что $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.

Предложение. Для всякой *невырожденной* квадратной $n\times n$ матрицы $\mathbf{A}$ существует (единственная) обратная матрица $\mathbf{A}^{-1}$.

Предложение. Обозначим через a^{ij} элементы обратной матрицы $\mathbf{A}^{-1}$. Тогда $a^{ij} = \dfrac{(-1)^{i+j}\left|\mathbf{M}_{ji}\right|}{|\mathbf{A}|}$, где $\mathbf{M}_{ij}$ — матрица, получающаяся из $\mathbf{A}$ вычеркиванием i-й строки и j-го столбца.

Примеры.

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix},$$

$$\begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{pmatrix}^{-1} = \begin{pmatrix} \lambda_1^{-1} & 0 & \dots & 0 \\ 0 & \lambda_2^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n^{-1} \end{pmatrix}.$$

Предложение. Свойства обратной матрицы:

$\left|\mathbf{A}^{-1}\right| = |\mathbf{A}|^{-1}$,

$(\mathbf{A}^{-1})^{-1} = \mathbf{A}$,

$(\mathbf{A}^{-1})^{T} = (\mathbf{A}^{T})^{-1}$,

если существуют $\mathbf{A}^{-1}$ и $\mathbf{B}^{-1}$, то $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

(Отметим, что в последней формуле все матрицы *квадратные* и невырожденные.)

11. Системы линейных уравнений.

Систему n линейных уравнений с n неизвестными

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1, \\ \dots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = b_n. \end{cases}$$

удобно записать в матричном виде:

$\mathbf{Ax}=\mathbf{b}$, где $\mathbf{A} = (a_{ij})$ — квадратная $n\times n$ матрица, $x = (x_1,\dots,x_n)^T$ и $b = (b_1,\dots,b_n)^T$ — векторы-столбцы.

Предложение. Если матрица $\mathbf{A}$ невырожденная, то система имеет единственное решение: $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

Предложение. Однородная система $\mathbf{Ax} = \mathbf{0}$ имеет ненулевое решение тогда и только тогда, когда матрица $\mathbf{A}$ вырожденная: $|\mathbf{A}| = 0$.

12. Собственные числа и векторы.

Пусть $\mathbf{A}$ матрица $n\times n$.

Определение. Любой ненулевой вектор $\mathbf{x}$ — некоторый вектор, для которого выполняется равенство $\mathbf{Ax}=\lambda\mathbf{x}$, где λ — некоторое число, которое называется собственным значением матрицы $\mathbf{A}$, а $\mathbf{x}$ — называется собственным вектором матрицы $\mathbf{A}$.

Тогда $\mathbf{Ax} = \lambda\mathbf{x} \Rightarrow (\mathbf{A}-\lambda\mathbf{I})\mathbf{x} = 0$, где $\mathbf{I}$ — единичная матрица.

Решение последнего уравнения для заданного λ позволяет получить соответствующий собственный вектор матрицы $\mathbf{A}$.

Определение. Многочлен $\det(\mathbf{A}-\lambda\mathbf{I})$ называется характеристическим многочленом матрицы $\mathbf{A}$, а уравнение $\det(\mathbf{A}-\lambda\mathbf{I}) = 0$ характеристическим уравнением для матрицы $\mathbf{A}$.

Решением уравнения $\det(\mathbf{A}-\lambda\mathbf{I}) = 0$ являются собственные значения матрицы $\mathbf{A}$ (т.е. λ_i, i,=1,2,...,n.).

Пример.

$$\mathbf{A} = \begin{pmatrix} 1 & 4 \\ 1 & 1 \end{pmatrix},$$

$$\det(\mathbf{A}-\lambda\mathbf{I}) = \begin{vmatrix} 1-\lambda & 4 \\ 1 & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - 4 = \lambda^2 - 2\lambda - 3 = 0, \quad \lambda_1 = -1, \lambda_2 = 3.$$

При $\lambda_1 = -1$, например, получаем $\begin{cases} (1-(-1))x_1 + 4x_2 = 0, \\ x_1 + (1-(-1))x_2 = 0, \end{cases} \Rightarrow$

$\mathbf{x} = \begin{pmatrix} -2c \\ c \end{pmatrix}$, где c — произвольное число, не ноль.

Предложение. Разным собственным числам соответствуют линейно *независимые* собственные векторы.

Предложение. Пусть характеристический многочлен матрицы $\mathbf{A}$ имеет n *различных вещественных* корней. Тогда мат-

рица $\mathbf{A}$ может быть представлена в виде $\mathbf{A} = \mathbf{C}^{-1}\Lambda\mathbf{C}$, где матрица Λ — *диагональная*, а матрица $\mathbf{C}$ — невырожденная.

13. Симметричные матрицы.

Определение. Матрица $\mathbf{A}$ называется *симметричной*, если $\mathbf{A}^T = \mathbf{A}$.

Предложение. Для любой матрицы $\mathbf{A}$ матрица $\mathbf{A}^T\mathbf{A}$ — симметричная.

Предложение. Симметричная $n \times n$ матрица $\mathbf{A}$ имеет n собственных чисел (некоторые из них могут совпадать), которым соответствуют n собственных векторов $\mathbf{c}_1, \ldots, \mathbf{c}_n$, которые могут быть выбраны *попарно ортогональными*. (Собственные векторы, соответствующие *разным* собственным значениям симметричной матрицы, всегда ортогональны.)

Более того, поскольку собственный вектор определяется с точностью до коэффициента пропорциональности, то можно нормировать собственные векторы $\{\mathbf{c}_i\}$ так, что они будут *ортонормированной системой*, т. е. попарно ортогональны и единичной длины.

Тогда матрица $\mathbf{A}$ приводится к диагональному виду при помощи матрицы $\mathbf{O}$, столбцы которой являются векторами $\mathbf{c}_i$: $\Lambda = \mathbf{O}^{-1}\mathbf{A}\mathbf{O}$, где на диагонали матрицы Λ стоят собственные числа матрицы $\mathbf{A}$.

Определение. Матрица, столбцы которой составляют ортонормированную систему векторов, называется *ортогональной*.

Предложение. Ортогональная матрица удовлетворяет соотношению: $\mathbf{O}^T\mathbf{O} = \mathbf{I}$.

Предложение. Если $\mathbf{O}$ — ортогональная матрица, то $\mathbf{O}^T = \mathbf{O}^{-1}$.

Предложение. Ортогональная матрица имеет определитель, равный $+1$ или -1.

Предложение. В ортогональной матрице строки также образуют ортонормированную систему векторов.

Предложение. Симметричная матрица $\mathbf{A}$ может быть приведена к диагональному виду при помощи ортогонального преобразования $\mathbf{O}$: $\mathbf{O}^T\mathbf{A}\mathbf{O} = \Lambda$.

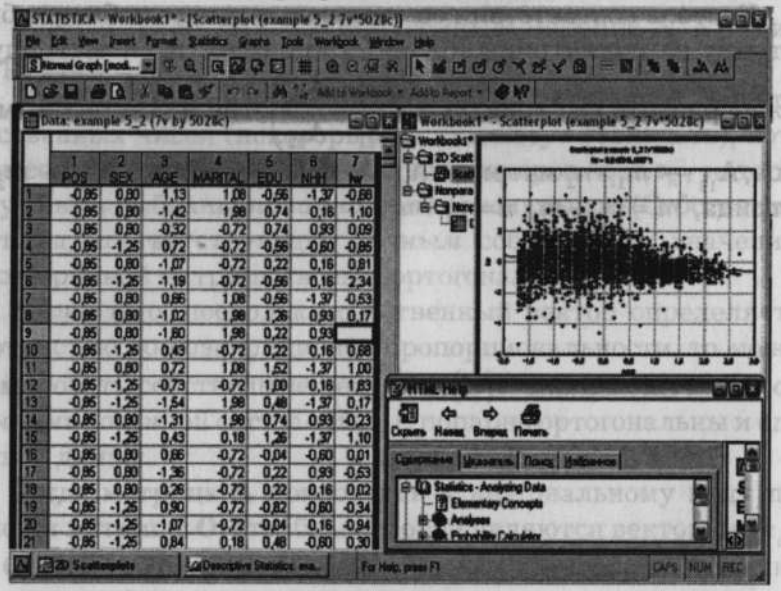
Предложение. Последнее соотношение можно записать в виде разложения симметричной матрицы $\mathbf{A}$ на ортогональную и диагональную: $\mathbf{A} = \mathbf{O}\Lambda\mathbf{O}^T$, где на диагонали матрицы Λ стоят собственные числа матрицы $\mathbf{A}$.

14. Блочные матрицы.

Часто, в соответствии со смыслом задачи, удобно разбить матрицу на подматрицы (блоки). Например, $m \times n$ матрицу A можно разбить на блоки: $\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$, где $\mathbf{A}_{11}$ — $m_1 \times n_1$ матрица, $\mathbf{A}_{12}$ — $m_1 \times n_2$ матрица, $\mathbf{A}_{21}$ — $m_2 \times n_1$ матрица, $\mathbf{A}_{22}$ — $m_2 \times n^2$ матрица, $m = m_1 + m_2$, $n = n_1 + n_2$.

Введение в Statistica

Запуск программы осуществляется из главного меню «Пуск». Рабочее окно программы имеет вид:



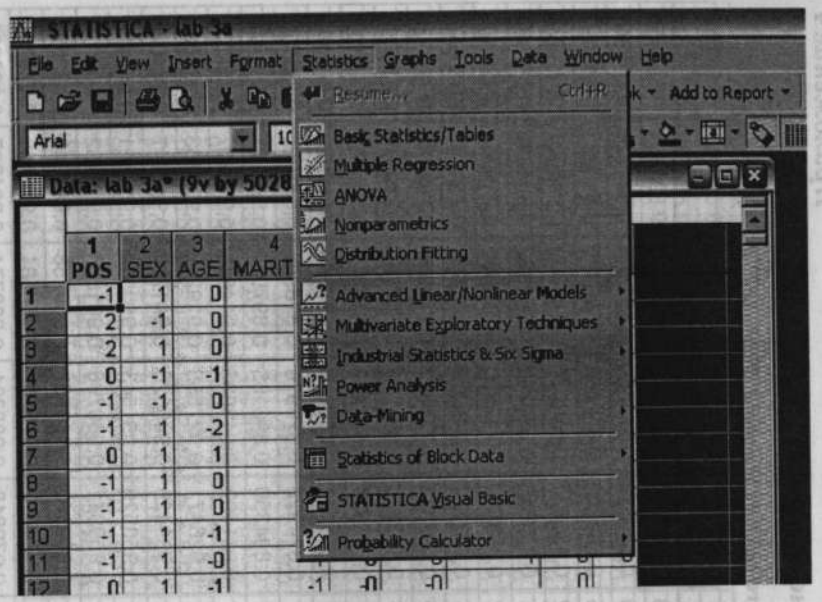
Структура окна программы похожа на стандартные приложения Windows и состоит из заголовка окна, главного меню, панелей инструментов, рабочей области, в которой отображаются текущие загруженные данные.

Исходные данные и результаты расчетов оформлены в виде таблиц. Исходные данные по столбцам содержат переменные (variables), а строки-случаи (cases).

Запуск основных модулей анализа данных осуществляется с помощью опций главного меню Statistics.

Структура диалога при выборе того или иного типа анализа (модуля программы) имеет общие черты.

Для открытого файла данных выбирают тип анализа, затем выбирают переменные и метод, наконец, конкретную вычислительную процедуру и задают ее параметры.



После запуска процедуры и получения результатов с помощью вкладок возможно получение выводов в графической или табличной формах.

Подробнее о возможностях программы можно ознакомиться с помощью справки. Также рекомендуется различных лет издания книга — *Боровиков В.П.* Популярное введение в программу Statistica. Краткое руководство по пакету имеется на сайте www.exponenta.ru в сети Интернет.

Статистические таблицы

Функция стандартного нормального распределения $\Phi(x) = \int_{-\infty}^{x} e^{-t^2/2} dt$.

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,500000	0,503989	0,507978	0,511967	0,515953	0,519939	0,523922	0,527903	0,531881	0,535856
0,10	0,539828	0,543795	0,547758	0,551717	0,555670	0,559618	0,563559	0,567495	0,571424	0,575345
0,20	0,579260	0,583166	0,587064	0,590954	0,594835	0,598706	0,602568	0,606420	0,610261	0,614092
0,30	0,617911	0,621719	0,625516	0,629300	0,633072	0,636831	0,640576	0,644309	0,648027	0,651732
0,40	0,655422	0,659097	0,662757	0,666402	0,670031	0,673645	0,677242	0,680822	0,684386	0,687933
0,50	0,691462	0,694974	0,698468	0,701944	0,705402	0,708840	0,712260	0,715661	0,719043	0,722405
0,60	0,725747	0,729069	0,732371	0,735653	0,738914	0,742154	0,745373	0,748571	0,751748	0,754903
0,70	0,758036	0,761148	0,764238	0,767305	0,770350	0,773373	0,776373	0,779350	0,782305	0,785236
0,80	0,788145	0,791030	0,793892	0,796731	0,799546	0,802338	0,805106	0,807850	0,810570	0,813267
0,90	0,815940	0,818589	0,821214	0,823814	0,826391	0,828944	0,831472	0,833977	0,836457	0,838913
1,00	0,841345	0,843752	0,846136	0,848495	0,850830	0,853141	0,855428	0,857690	0,859929	0,862143
1,10	0,864334	0,866500	0,868643	0,870762	0,872857	0,874928	0,876976	0,878999	0,881000	0,882977
1,20	0,884930	0,886860	0,888767	0,890651	0,892512	0,894350	0,896165	0,897958	0,899727	0,901475
1,30	0,903199	0,904902	0,906582	0,908241	0,909877	0,911492	0,913085	0,914656	0,916207	0,917736
1,40	0,919243	0,920730	0,922196	0,923641	0,925066	0,926471	0,927855	0,929219	0,930563	0,931888
1,50	0,933193	0,934478	0,935744	0,936992	0,938220	0,939429	0,940620	0,941792	0,942947	0,944083
1,60	0,945201	0,946301	0,947384	0,948449	0,949497	0,950529	0,951543	0,952540	0,953521	0,954486
1,70	0,955435	0,956367	0,957284	0,958185	0,959071	0,959941	0,960796	0,961636	0,962462	0,963273
1,80	0,964070	0,964852	0,965621	0,966375	0,967116	0,967843	0,968557	0,969258	0,969946	0,970621
1,90	0,971284	0,971933	0,972571	0,973197	0,973810	0,974412	0,975002	0,975581	0,976148	0,976705
2,00	0,977250	0,977784	0,978308	0,978822	0,979325	0,979818	0,980301	0,980774	0,981237	0,981691

Окончание таблицы значений функции стандартного нормального распределения

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
2,10	0,982136	0,982571	0,982997	0,983414	0,983823	0,984222	0,984614	0,984997	0,985371	0,985738
2,20	0,986097	0,986447	0,986791	0,987126	0,987455	0,987776	0,988089	0,988396	0,988696	0,988989
2,30	0,989276	0,989556	0,989830	0,990097	0,990358	0,990613	0,990863	0,991106	0,991344	0,991576
2,40	0,991802	0,992024	0,992240	0,992451	0,992656	0,992857	0,993053	0,993244	0,993431	0,993613
2,50	0,993790	0,993963	0,994132	0,994297	0,994457	0,994614	0,994766	0,994915	0,995060	0,995201
2,60	0,995339	0,995473	0,995603	0,995731	0,995855	0,995975	0,996093	0,996207	0,996319	0,996427
2,70	0,996533	0,996636	0,996736	0,996833	0,996928	0,997020	0,997110	0,997197	0,997282	0,997365
2,80	0,997445	0,997523	0,997599	0,997673	0,997744	0,997814	0,997882	0,997948	0,998012	0,998074
2,90	0,998134	0,998193	0,998250	0,998305	0,998359	0,998411	0,998462	0,998511	0,998559	0,998605
3,00	0,998650	0,998694	0,998736	0,998777	0,998817	0,998856	0,998893	0,998930	0,998965	0,998999
3,10	0,999032	0,999064	0,999096	0,999126	0,999155	0,999184	0,999211	0,999238	0,999264	0,999289
3,20	0,999313	0,999336	0,999359	0,999381	0,999402	0,999423	0,999443	0,999462	0,999481	0,999499
3,30	0,999517	0,999533	0,999550	0,999566	0,999581	0,999596	0,999610	0,999624	0,999638	0,999650
3,40	0,999663	0,999675	0,999687	0,999698	0,999709	0,999720	0,999730	0,999740	0,999749	0,999758
3,50	0,999767	0,999776	0,999784	0,999792	0,999800	0,999807	0,999815	0,999821	0,999828	0,999835
3,60	0,999841	0,999847	0,999853	0,999858	0,999864	0,999869	0,999874	0,999879	0,999883	0,999888
3,70	0,999892	0,999896	0,999900	0,999904	0,999908	0,999912	0,999915	0,999918	0,999922	0,999925
3,80	0,999928	0,999930	0,999933	0,999936	0,999938	0,999941	0,999943	0,999946	0,999948	0,999950
3,90	0,999952	0,999954	0,999956	0,999958	0,999959	0,999961	0,999963	0,999964	0,999966	0,999967
4,00	0,999968	0,999970	0,999971	0,999972	0,999973	0,999974	0,999975	0,999976	0,999977	0,999978

Например, $\Phi(1{,}54) = 0{,}938220$.

Двусторонние квантили t - распределения Стьюдента

Число степеней свободы	Уровень значимости						
	0,1	0,05	0,025	0,02	0,01	0,005	0,001
1	6,314	12,706	25,452	31,821	63,657	127,300	636,600
2	2,920	4,303	6,205	6,965	9,925	14,089	31,598
3	2,353	3,182	4,177	4,541	5,841	7,453	12,941
4	2,132	2,776	3,495	3,747	4,604	5,597	8,610
5	2,015	2,571	3,163	3,365	4,032	4,773	6,859
6	1,943	2,447	2,969	3,143	3,707	4,317	5,959
7	1,895	2,365	2,841	2,998	3,499	4,029	5,405
8	1,860	2,306	2,752	2,896	3,355	3,833	5,041
9	1,833	2,262	2,685	2,821	3,250	3,690	4,781
10	1,812	2,228	2,634	2,764	3,169	3,581	4,587
11	1,796	2,201	2,593	2,718	3,106	3,497	4,437
12	1,782	2,179	2,560	2,681	3,055	3,428	4,318
13	1,771	2,160	2,533	2,650	3,012	3,372	4,221
14	1,761	2,145	2,510	2,624	2,977	3,326	4,140
15	1,753	2,131	2,490	2,602	2,947	3,286	4,073
16	1,746	2,120	2,473	2,583	2,921	3,252	4,015
17	1,740	2,110	2,458	2,567	2,898	3,222	3,965
18	1,734	2,101	2,445	2,552	2,878	3,193	3,922
19	1,729	2,093	2,433	2,539	2,861	3,174	3,883
20	1,725	2,086	2,423	2,528	2,845	3,153	3,849
21	1,721	2,080	2,414	2,518	2,831	3,135	3,819
22	1,717	2,074	2,405	2,508	2,819	3,119	3,792
23	1,714	2,069	2,398	2,500	2,807	3,104	3,768
24	1,711	2,064	2,391	2,492	2,797	3,092	3,745
25	1,708	2,060	2,385	2,485	2,787	3,078	3,725
26	1,706	2,056	2,379	2,479	2,779	3,067	3,707
27	1,703	2,052	2,373	2,473	2,771	3,057	3,689
28	1,701	2,048	2,369	2,467	2,763	3,047	3,674
29	1,699	2,045	2,364	2,462	2,756	3,038	3,660
30	1,697	2,042	2,360	2,457	2,750	3,030	3,646
∞	1,645	1,960	2,241	2,326	2,576	2,807	3,291

Квантили распределения χ^2

Число степеней свободы	Уровень значимости					
	0,50	0,30	0,20	0,10	0,05	0,01
1	0,455	1,074	1,642	2,706	3,841	6,635
2	1,386	2,408	3,219	4,605	5,991	9,210
3	2,366	3,665	4,642	6,251	7,815	11,345
4	3,357	4,878	5,989	7,779	9,488	13,277
5	4,351	6,064	7,289	9,236	11,071	15,086
6	5,348	7,231	8,558	10,645	12,592	16,812
7	6,346	8,383	9,803	12,017	14,067	18,475
8	7,344	9,524	11,030	13,362	15,507	20,090
9	8,343	10,656	12,242	14,684	16,919	21,666
10	9,342	11,781	13,442	15,987	18,307	23,209
11	10,341	12,899	14,631	17,275	19,675	24,725
12	11,340	14,011	15,812	18,549	21,026	26,217
13	12,340	15,119	16,985	19,812	22,362	27,688
14	13,339	16,222	18,151	21,064	23,685	29,141
15	14,339	17,322	19,311	22,307	24,996	30,578
16	15,339	18,418	20,465	23,542	26,296	32,000
17	16,338	19,511	21,615	24,769	27,587	33,409
18	17,338	20,601	22,760	25,989	28,869	34,805
19	18,338	21,689	23,900	27,204	30,144	36,191
20	19,337	22,775	25,038	28,412	31,410	37,566
21	20,337	23,858	26,171	29,615	32,671	38,932
22	21,337	24,939	27,301	30,813	33,924	40,289
23	22,337	26,018	28,429	32,007	35,172	41,638
24	23,337	27,096	29,553	33,196	36,415	42,980
25	24,337	28,172	30,675	34,382	37,652	44,314
26	25,336	29,246	31,795	35,563	38,885	45,642
27	26,336	30,319	32,912	36,741	40,113	46,963
28	27,336	31,391	34,027	37,916	41,337	48,278
29	28,336	32,461	35,139	39,087	42,557	49,588
30	29,336	33,530	36,250	40,256	43,773	50,892

Распределение Фишера $F(n_1, n_2)$, $\varepsilon = 0{,}05$

	n_2																		
n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92

Окончание таблицы распределения Фишера $F(n_1, n_2)$, $\varepsilon = 0{,}05$

	n_2																		
n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

n_1 — число степеней свободы числителя, n_2 — число степеней свободы знаменателя

Распределение Фишера $F(n_1, n_2)$, $\varepsilon = 0{,}01$

	n_2																		
n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6107	6157	6209	6234	6260	6286	6313	6340	6366
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,48	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57

Окончание таблицы распределения Фишера $F(n_1, n_2)$, $\varepsilon = 0{,}01$

	n_2																		
n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

n_1 — число степеней свободы числителя, n_2 — число степеней свободы знаменателя

Учебное издание

Ниворожкина Л. И., Арженовский С. В.

Многомерные статистические методы
в экономике

Санитарно-эпидемиологическое заключение
№ 77.99.02.953.Д.004609.07.04 от 13.07.2004 г.

Подписано в печать 10.02.2009. Формат 60×84 1/16.
Печать офсетная. Бумага офсетная № 1. Печ. л. 14,0.
Тираж 1000 экз. Заказ №

Издательско-торговая корпорация «Дашков и К°»
129347, Москва, Ярославское шоссе, д. 142, к. 732.
Для писем: 129347, Москва, п/о И-347
Тел./факс: (499) 182-01-58, 182-11-79, 183-93-01
E-mail: sales@dashkov.ru — отдел продаж
office@dashkov.ru — офис;
http://www.dashkov.ru

Отпечатано в соответствии с качеством предоставленных диапозитивов
в ФГУП «Производственно-издательский комбинат ВИНИТИ»,
140010, г. Люберцы Московской обл., Октябрьский пр-т, 403. Тел.: 554-21-86